LLM 기반 AI 에이전트의 신뢰성 있는 복합 명령 처리를 위한 다단계 추론 기법

이남경, 김혜린, 김희원, 이건희*
에이치디씨랩스

{nk_lee, helen6339, ive2go, Gunhee_Lee}@hdc-labs.com

Multi-Step Reasoning for Reliable Composite Command Handling in LLM-Based AI Agents

Namkyeong LEE, Hye-Lynn Kim, Heewon Kim, Gunhee Lee*
HDC LABS.

요 약

본 논문은 LLM(gemma-2-9b-it-AWQ-INT4)을 활용한 AI Agent 시스템에서 복합 제어 명령을 처리하는 추론 구조에 따른 성능 차이를 분석한다. 기존의 단일 추론 기반 기능 분류 방식과, 본 연구에서 제안하는 다단계 추론 기반 기능별 질의 분해 방식을 실제 스마트홈 운영 로그 데이터를 통해 비교하였다. 실험 결과, 제안 방식은 기능 분류의 재현율을 유지하면서도 잘못된 기능 호출을 크게 감소시켜, 복합 제어 환경에서 보다 안정적인 AI Agent 동작을 가능하게 함을 확인하였다.

Ⅰ. 서 론

최근 LLM 의 급격한 성능 향상으로 자연어 기반 AI Agent 는 스마트홈, 로봇 제어, 산업 자동화 등 다양한 영역에서 핵심 인터페이스로 자리 잡고 있다. 사용자는 더 이상 정형화된 명령어를 학습할 필요 없이, 일상적인 자연어로 복합적인 제어 의도를 표현할 수 있으며, AI Agent 는 이를 해석하여 여러 제어 기능을 순차적으로 실행한다.

특히 스마트홈 환경에서는 “불 켜고 난방 켜줘”, “외출 모드 설정하고 엘리베이터 불러줘”와 같이 하나의 발화에 다수의 제어 기능이 동시에 포함되는 경우가 빈번하다. 이러한 복합 제어 명령은 사용자 경험 측면에서는 매우 직관적이지만, AI Agent 입장에서는 각 기능을 정확히 분리·매핑해야 하는 높은 난이도의 추론 문제를 내포한다.

복합 명령을 정확히 처리하지 못할 경우, 불필요한 기능이 실행되거나 의도되지 않은 제어가 발생할 수 있다. 예를 들어 “불은 끄고 난방은 유지해줘” 라는 명령에서 난방을 끄는 오동작이 발생한다면, 이는 단순한 인식 오류를 넘어 실제 사용자 신뢰도에 직접적인 영향을 미치는 문제로 이어진다. 따라서 복합 제어 명령 처리에서의 정확성 뿐만 아니라 안정성은 AI Agent 시스템 설계의 핵심 요구사항이라 할 수 있다.

기존의 많은 LLM 기반 AI Agent 시스템에서는 복합 명령을 단일 추론 과정으로 처리하는 방식을 채택해 왔다.[1] 이 방식에서는 LLM 이 전체 명령을 한 번에 해석한 뒤, 각 기능에 대한 판단 사유(reason)와 기능 인덱스(index)를 동시에 출력한다. 이러한 접근은 구현이 단순하고 추론 결과를 사람이 해석하기 쉽다는 장점이 있다.

그러나 단일 추론 방식은 복합 명령의 길이와 포함 기능 수가 증가할수록 추론 범위가 급격히 확장되며, 그 결과 불필요한 기능 예측이나 기능간 혼선이 발생할 가능성이 커진다. 특히 LLM 이 “설명 가능한 추론(reason)”을 함께 생성하도록 요구받을 경우, 실제 실행과 무관한 추론 토큰이 증가하면서 오히려 제어 안정성을 저해하는 사례가 관찰된다.

이에 본 연구에서는 복합 제어 명령을 단일 추론 문제가 아닌, 다단계 추론 문제로 재정의한다. 즉, 복합 명령을 한번에 해석하기보다는 기능 단위로 분리하여 단계적으로 처리함으로써, 각 기능이 자신에게 필요한 정보만을 처리하도록 하는 구조를 제안한다. 본 논문은 이러한 다단계 추론 기반 접근이 기존 단일 추론 방식 대비 어떤 성능적 이점을 가지는지를 실제 산업 환경 데이터를 통해 실증적으로 분석한다.

Ⅱ. 본론

1. 데이터셋 구축 및 실험 환경

본 연구는 실제 스마트홈 AI Agent 시스템에서 수집된 사용자 음성 및 텍스트 기반 제어 명령 로그를 활용하여 실험을 수행하였다. 해당 시스템은 상용 환경에서 운용중이며, 다양한 연령대와 사용 패턴을 가진 사용자로부터 자연어 명령이 지속적으로 수집된다.

실험에 사용된 데이터셋은 사용자 자연어 제어 명령, 명령에 포함된 실제 의도된 기능 목록 (전문가 수작업 라벨링), AI Agent 의 실제 기능 실행 로그로 구성된다.

복합 제어 명령은 하나 이상의 기능(조명, 난방, 가스, 환기, 외출 모드, 엘리베이터 호출 등)을 동시에 포함하는 경우로 정의하였다. 각 명령에 대해 도메인 전문가가 기능 단위의 정답 레이블을 부여하여, 정량적인 성능 평가가 가능하도록 구성하였다.

2. 복합 제어 명령 처리 전략

2.1 단일 추론 기반 분류 방식

단일 추론 기반 방식에서는 LLM 이 하나의 프롬프트를 통해 전체 명령을 분석하고, 명령에 포함된 기능들의 인덱스와 판단 사유(reason)를 동시에 출력한다. 이 방식은 직관적인 구조를 가지며, LLM 의 추론 능력을 최대한 활용할 수 있다는 장점이 있다.

그러나 실험 과정에서 이 방식은 복합 명령의 길이가 증가하거나 기능 수가 많아질수록 불필요한 기능을 함께 예측하는 경향을 보였다. 특히 “확인”, “유지”, “알려줘”와 같은 비제어성 표현이 포함될 경우, 실제 실행이 필요 없는 기능이 제어 대상으로 분류되는 문제가 관찰되었다.

2.2 다단계 추론 기반 기능별 질의 분해 방식 (제안 방법)

제안하는 방식은 복합 제어 명령을 두 개의 추론 단계로 분리한다.

첫 번째 단계에서는 LLM 을 이용해 명령에 포함된 기능 후보를 거칠게 분류한다. 이 단계는 높은 재현율을 목표로 하여, 실제 필요한 기능을 놓치지 않는 것을 우선시한다.

두 번째 단계에서는 각 기능에 대해 독립적인 LLM 질의를 수행한다. LLM 은 특정 기능 설명(feature description)과 사용자 명령을 함께 입력받아, 해당 기능과 직접적으로 관련된 문장만을 추출하거나, 관련이 없는 경우 “NONE”을 반환한다. 이 과정을 통해 기능 간 간섭을 제거하고, 실제 실행 가능한 기능만을 최종적으로 선택한다.

이러한 구조는 LLM 이 한 번에 모든 추론을 수행하도록 강제하는 대신, 추론 범위를 단계적으로 제한함으로써 안정적인 제어 판단을 가능하게 한다.[2]

3. 성능 평가 지표 및 실험 결과

복합 제어 명령 처리에서 중요한 성능 요소는 사용자가 의도한 기능을 빠짐없이 인식하는 것뿐만 아니라, 의도하지 않은 기능이 실행되지 않도록 억제하는 것이다. 이에 본 연구에서는 기능 분류 성능을 정밀도(Precision), 재현율(Recall), F1-score 및 잘못된 기능 호출 횟수(False Positive, FP)를 기준으로 평가하였다.

표 1. 단일 추론 방식과 다단계 추론 방식의 복합 명령 처리 성능 비교

	FP	Precision	F1	Recall
단일 추론방식	110	0.59	0.74	0.99
다단계 추론방식	6	0.96	0.98	0.98

실험 결과, 단일 추론 기반 방식은 재현율(Recall) 측면에서는 높은 성능을 보였으나, 복합 명령 내 비관련 기능까지 함께 선택하는 경향으로 인해 잘못된 기능 호출(False Positive)이 다수 발생하였다. 이로 인해 정밀도(Precision)가 0.59 수준에 머물렀으며, 전체 분류 성능을 나타내는 F1-score 는 0.74 로 나타났다.

반면, 제안하는 다단계 추론 기반 방식은 재현율을 거의 동일한 수준으로 유지하면서도 잘못된 기능 호출을 크게 감소시켜 정밀도를 0.96 까지 향상시켰으며, 그 결과 F1-score 가 0.98 로 유의미하게 개선되었다. 이는 제안 방식이 복합 제어 환경에서 불필요한 기능 실행을 효과적으로 억제함으로써, 보다 안정적이고 신뢰성 높은 제어를 가능하게 함을 보여준다.

Ⅲ. 결론

본 논문에서는 LLM 기반 AI Agent 시스템에서 복합 제어 명령을 처리하는 추론 구조의 차이가 제어 성능과 시스템 안정성에 미치는 영향을 분석하였다. 특히, 기존의 단일 추론 기반 기능 분류 방식과 본 연구에서 제안한 다단계 추론 기반 기능별 질의 분해 방식을 실제 스마트홈 운영 로그 데이터를 기반으로 비교·평가하였다.

실험 결과, 단일 추론 기반 방식은 재현율 측면에서는 높은 성능을 보였으나, 복합 명령 내 비관련 기능까지 함께 선택하는 경향으로 인해 잘못된 기능 호출이 빈번하게 발생하였다. 반면, 제안한 다단계 추론 기반 방식은 재현율을 거의 동일한 수준으로 유지하면서도 잘못된 기능 호출을 현저히 감소시켜, 정밀도와 F1-score 측면에서 유의미한 성능 향상을 달성하였다. 이는 복합 제어 환경에서 단순한 기능 인식 정확도보다, 불필요한 기능 실행을 억제하는 구조적 설계가 시스템 신뢰성에 더 큰 영향을 미친다는 점을 실증적으로 보여준다.

본 연구에서 제안한 다단계 추론 기반 접근은 스마트홈 환경뿐만 아니라, 로봇 제어, 산업 자동화, 멀티툴 AI Agent 등 복합적인 제어 의도가 요구되는 다양한 응용 분야로 확장 가능하다. 향후 연구에서는 추론 단계 수 증가에 따른 추론 비용과 지연(latency) 간의 trade-off 를 분석하고, 기능 후보 선택 단계의 자동 최적화 및 경량화 기법을 도입함으로써 실시간 제어 환경에서의 적용 가능성을 더욱 확대할 예정이다.

참 고 문 헌

[1] Li, S. et al., “HomeBench: Evaluating LLMs in Smart Homes with Valid Instruction-Device Call Data,” arXiv, 2025.

[2] Wei, J. et al., “Multi-Step Reasoning with Large Language Models: A Survey,” Commun. ACM, 2025.