멀티에이전트 언어 모델 시스템의 자동 평가에 대한 실험적 분석

이채은, *유 준

가천대학교

khce03@gachon.ac.kr, *joon.yoo@gachon.ac.kr

An Experimental Analysis of Automatic Evaluation for Multi-Agent Language Model Systems

Chaeeun Lee, *Joon Yoo

Gachon Univ.

요 약

본 논문에서는 CAMEL 기반 멀티에이전트 시스템과 단일 에이전트 시스템의 응답 품질을 AlpacaEval 벤치마크를 통해 비교·평가하였다. 소규모 실험에서는 멀티에이전트 시스템이 단일 에이전트 대비 상대적으로 높은 성능을 보였으나, 대규모 평가에서는 선호도 기반 지표에서 성능 향상이 제한적이거나 오히려 감소하는 경향이 관찰되었다. 이러한 결과를 통해 멀티에이전트 추론의 잠재적 장점과 함께, 자동 선호도 기반 평가 지표가 가지는 한계를 분석하고 향후 개선 방향을 제시한다.

Ⅰ. 서 론

최근 대규모 언어 모델(Large Language Model, LLM)을 기반으로 한 멀티에이전트 시스템은 복잡한 문제 해결과 추론 능력 향상을 목적으로 활발히 연구되고 있다. 특히 CAMEL(Communicative Agents for “Mind” Exploration of Large Language Model Society)[1]과 같은 프레임워크는 여러 에이전트 간의 협업을 통해 단일 에이전트 대비 보다 정교하고 일관된 응답을 생성하는 것을 목표로 한다.

그러나 이러한 멀티에이전트 구조가 실제 자동 평가 벤치마크에서 일관된 성능 향상을 보이는지에 대해서는 충분한 실험적 검증이 이루어지지 않았다. 기존 연구에서는 주로 정성적 분석이나 사례 중심의 평가가 수행되어, 협업 구조의 효과를 객관적인 수치로 비교하기에는 한계가 존재한다.

이에 본 연구에서는 CAMEL 기반 멀티에이전트 시스템과 단일 에이전트 시스템을 AlpacaEval[2]을 통해 비교 평가함으로써, 멀티에이전트 협업의 효과와 자동 평가 지표의 한계를 체계적으로 분석하고자 한다.

Ⅱ. 본론

2.1 멀티 에이전트 구조

CAMEL 은 두 개 이상의 에이전트에게 명확한 역할(role)을 부여하고, 역할 간 대화를 통해 공동의 목표를 달성하도록 설계된 멀티에이전트 프레임워크이다. 각 에이전트는 자신의 역할에 부합하는 관점에서 응답을 생성하며, 상호 피드백 과정을 통해 결과를 점진적으로 개선한다. 이러한 구조는 인간 협업 과정과 유사한 추론 방식을 모사한다는 점에서 의미가 있다.

그러나 CAMEL 논문에서 제시된 평가는 주로 정성적 분석에 머무르고 있어, 멀티에이전트 협업이 실제 성능 향상에 얼마나 기여하는지를 수치적으로 판단하기 어렵다. 이에 따라 멀티에이전트 구조의 효과를 보다 체계적으로 검증할 수 있는 정량적 평가 방법이 요구된다.

2.2 Pairwise Preference Evaluation

AlpacaEval 은 LLM 출력 평가의 주관성을 줄이기 위해 제안된 자동 평가 프레임워크이다 [2]. 이 방법은 단일 출력에 절대적인 점수를 부여하는 방식이 아니라, 동일한 입력에 대해 생성된 두 개의 출력을 직접 비교하여 어느 쪽이 더 우수한지를 판단하는 pairwise preference evaluation 방식을 사용한다.

Dubois 등은 이러한 비교 기반 평가 방식이 인간 평가자 간 일관성이 높으며, LLM 을 평가자로 활용하였을 때에도 인간 선호와 높은 상관관계를 보인다는 점을 실험적으로 입증하였다. 본 연구에서는 이 평가 방식을 활용하여 CAMEL 멀티에이전트 시스템과

단일 에이전트 시스템의 출력을 동일한 조건에서 비교하였다.

구체적으로, 동일한 instruction 입력에 대해 다음 두 시스템을 평가하였다.

- 멀티에이전트 시스템: 여러 에이전트가 상호작용하며 응답을 생성
- 단일 에이전트 시스템: 단일 언어 모델의 출력을 사용

생성된 결과는 AlpacaEval 의 pairwise preference 평가를 통해 비교되었으며, 시스템 구조 차이가 응답의 길이, 구조적 특성, 그리고 선호도 점수에 미치는 영향을 분석하였다.

2.3 실험

실험은 CAMEL GitHub 에 공개된 *camel_code* 데이터셋을 활용하였으며, 단일 에이전트 시스템에는 GPT-4o-mini 모델을 사용하였다.

표 1. AlpacaEval 기반 CAMEL 멀티에이전트 시스템의 소규모 평가 결과

Model	*Win rate (%)*	*Length-controlled win rate (%)*	*n*
CAMEL (Multi-agent)	8.60	14.58	28

표 2. AlpacaEval 기반 CAMEL 멀티에이전트 시스템의 대규모 평가

Model	*Win rate (%)*	*Length-controlled win rate (%)*	*n*
CAMEL (Multi-agent)	1.32	1.91	200

표 1 은 AlpacaEval 을 기반으로 28 개의 지시문에 대해 수행한 소규모 평가 결과를 나타낸다. 멀티에이전트 시스템은 단일 에이전트 대비 8.6%의 win rate 를 기록하였으며, 출력 길이를 통제한 경우 성능 향상이 관찰되었다.

표 2 에 제시된 200 개의 지시문에 대한 대규모 평가 결과에서 멀티에이전트 시스템의 win rate 는 1.32%로 나타났으며, length-controlled win rate 또한 제한적인 수준에 머물렀다.

분석결과 멀티에이전트 시스템은 장황하고 구조화된 응답을 생성하는 경향이 있으며, AlpacaEval 과 같은 선호도 기반 평가에서는 이러한 특성이 오히려 성능을 낮출 수 있다.

Ⅲ. 결론

본 연구에서는 CAMEL 기반 멀티에이전트 시스템과 단일 에이전트 시스템을 AlpacaEval 을 활용하여 비교 평가하였다. 실험 결과, 멀티에이전트 시스템은 소규모 평가에서는 일정 수준의 성능 향상을 보였으나, 대규모 실험에서의 win rate 는 전반적으로 높지 않은 것으로 나타났다. 이러한 결과는 멀티에이전트 협업 자체의 비효율성보다는, 선호도 기반 자동 평가 프레임워크가 멀티에이전트의 특성을 충분히 반영하지 못한 데에서 기인한 것으로 해석할 수 있다.

멀티에이전트 시스템은 문제를 단계적으로 분해하고, 다양한 관점에서의 논의를 거쳐 구조화된 응답을 생성하는 경향이 있다. 그러나 AlpacaEval 은 두 출력 중 어느 쪽이 더 “선호되는지”를 판단하는 비교 기반 평가 방식으로, 응답의 추론 과정이나 협업으로 인한 중간 reasoning 의 질을 직접적으로 고려하지 않는다. 이로 인해 멀티에이전트 시스템이 생성한 장황하고 구조적인 응답은 오히려 단일 에이전트의 간결한 응답에 비해 불리하게 평가될 가능성이 있다. 특히 대규모 평가로 확장될수록 이러한 경향이 누적되면서 win rate 의 개선 효과가 제한적으로 나타난 것으로 판단된다. 또한 AlpacaEval 에서 사용되는 평가 기준은 주로 최종 출력의 자연스러움과 즉각적인 선호도에 초점을 두고 있어, 멀티에이전트 협업이 제공하는 논리적 완결성이나 문제 해결 과정의 안정성과 같은 요소를 충분히 반영하기 어렵다. 이는 멀티에이전트 시스템이 지향하는 설계 목표와 평가 지표 간의 구조적 불일치를 시사한다.

향후 연구에서는 이러한 한계를 극복하기 위해, 멀티에이전트 시스템의 특성을 반영할 수 있는 평가 방식의 개선이 필요하다. 예를 들어, 최종 응답뿐만 아니라 추론 단계의 일관성, 역할 간 상호작용의 기여도, 혹은 reasoning 품질을 분리하여 평가하는 다차원적 벤치마크를 도입할 수 있다. 또한 멀티에이전트 출력의 장황함을 완화하고 핵심 정보 중심으로 요약하는 출력 구조 최적화 전략을 적용함으로써, 선호도 기반 평가 지표와의 정합성을 높이는 방향도 고려할 수 있다.

종합적으로, 본 연구의 결과는 멀티에이전트 시스템의 잠재적 장점이 자동 선호도 기반 평가 지표에서는 충분히 드러나지 않을 수 있음을 보여준다. 이는 멀티에이전트 추론의 한계라기보다는, 이를 평가하는 기준과 방법에 대한 재고의 필요성을 시사하며, 향후 멀티에이전트 LLM 연구에서 평가 방법론의 중요성을 강조하는 근거로 활용될 수 있다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2026 년도 SW 중심대학사업의 결과로 수행되었음. (2021-0-01389)

참 고 문 헌

[1] Li, G. et al., *CAMEL: Cooperative Autonomous Multi-Agent Learning*, arXiv preprint, 2023.

[2] Dubois, Y. et al., *AlpacaEval: An Automatic Evaluator of Instruction-Following Models*, arXiv preprint, 2023.