다중 언어 모델 기반 멀티 에이전트 챗봇 토론 프레임워크를 통한 환각 탐지 및 교정

송인혁[0], 김은호, 박상근

경희대학교 소프트웨어융합대학

thddlsgur0105@khu.ac.kr, taemin4u@khu.ac.kr, sk.park@khu.ac.kr

A Multi-Agent Debate Framework of Multiple Language Models for Hallucination Detection and Correction

Inhyuk Song[0], Eunnho Kim, Sangkeun Park

College of Software, Kyung Hee University

요 약

대형 언어 모델(LLM)을 활용한 정보 탐색 과정이 대중화되면서, 챗봇에서의 환각 문제가 대두되고 있다. 이러한 환각 문제를 해결하기 위한 여러 접근 방식의 사전 연구들이 진행되었으나, 단일 학습 데이터에 기반을 두고 있기 때문에 학습 데이터의 편향 문제를 완전히 해소하지 못했다는 한계가 존재한다. 본 연구는 멀티 언어 모델을 활용한 멀티 에이전트 챗봇 토론 프레임워크를 개발해 단일 모델 접근의 제약을 뛰어넘어, 보다 신뢰할 수 있는 LLM 기반 검색 서비스를 구현하고자 한다.

1. 서 론*

최근 ChatGPT, Gemini 등 대형 언어 모델(LLM) 기반 생성형 AI 서비스가 일반인 사이에서 급속히 확산되면서, 정보 탐색 수단으로 전통적 검색 엔진 대신 LLM 챗봇을 활용하는 사례가 늘고 있다. 디지털 마케팅 업체 Ignite Visibility에서 실시한 설문조사에 따르면, 170명 응답자 중 62%가 서비스 검색에 ChatGPT, Gemini 등의 LLM 기반 챗봇을 활용한다고 답했다[1]. 사용자는 검색 키워드를 고민할 필요 없이 자연어 대화만으로도 원하는 정보를 손쉽게 얻을 수 있으므로, LLM 기반 검색이 빠르게 대중화되고 있다.

그러나 생성형 AI를 통한 정보 검색에서는 '환각(hallucination)'이라는 치명적 문제가 동반된다. LLM의 환각이란 모델이 입력에 존재하지 않거나 사실과 다른 내용을 그럴듯하게 생성하는 현상을 의미하며[2], 여기에는 논리 오류·추론 오류·수학적 오류·근거 없는 조작 등 여러 유형이 포함된다[3]. 환각은 법률, 의료, 심지어 뉴스 허위 기사 작성 등 여러 분야에서 발생할 수 있다[2]. 이러한 환각 현상은 사용자에게 사실과 다른 정보를 제공해 오판을 유도할 수 있다는 점에서 심각한 문제를 초래할 수 있다.

환각 문제를 해결하기 위해 단일 생성 모델의 답변 내용을 활용한 환각 방지[4, 5], 환각 탐지[2, 6] 등의 연구가 수행됐다. 그럼에도 불구하고 발생하는 환각 현상 해결을 위해, 모델의 답변을 정확하게 개선하려는 다양한 연구가 수행됐다[7, 8, 9]. 하지만 이 연구들은 공통적으로 단일 학습 데이터를 사용한 한 모델만으로 설계되었으므로 학습 데이터 편향과 모델 특성에 따른 한계를 완전히 해소하기 어렵다.

본 논문에서는 위 한계를 극복하고자 서로 다른 학습 데이터로 사전 학습된 복수의 생성형 모델을 활용하는 멀티 챗봇 토론 프레임워크를 제안한다. 각 모델이 생성한 답변과 근거를 상호 비교·검증하는 토론을 통해 환각을 효과적으로 탐지하고, 모델 간 관점을 융합한 개선된 답변을 생성할 수 있음을 실험적으로 보인다. 이로써 단일 모델의 한계를 넘어 신뢰할 수 있는 LLM 기반 검색 서비스를 구현하고자 한다.

2. 관련 연구

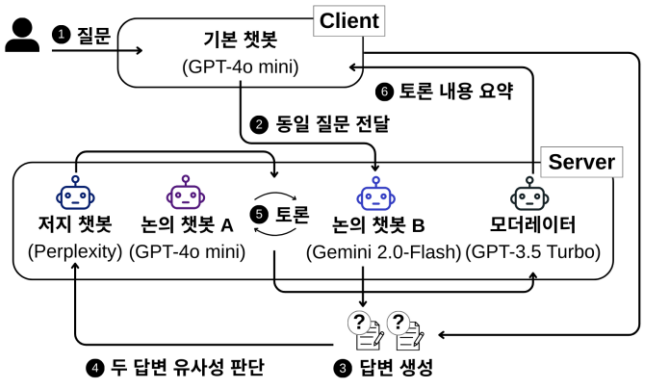
LLM 환각 문제를 해결하기 위해 환각 현상 방지를 위한 여러 연구들이 수행됐다. RAG를 활용해 외부 데이터셋에서 정보를 가져와 답변을 생성하는 방식[4], 혼합 대조 학습을 통해 모델에 환각 감지 정보를 학습시키는 방식이[5] 제안됐다. 그럼에도 발생하는 환각 현상을 탐지하기 위한 연구도 수행됐다. 한 모델이 생성한 다수의 답변 간 일관성을 분석해 환각 여부를 판단하는 방법과[6], 답변의 다양성을 기준으로 환각을 판별하는 접근법이[2] 제안됐다.

환각 탐지 후 답변 개선을 위한 시도들도 있었다. 모델에 환각 발생 근거를 추가 학습시켜 정확도를 높이거나[7], 답변을 외부 데이터셋에서 재 검증 및 보강하거나[8], 서로 다른 답변을 모델 간 토론 형태로 비교·검증하는 방식[9] 등이 제안됐다. 그러나 이들 연구는 단일 모델 기반으로 설계되어 학습 데이터 편향 문제를 완전히 해소하기 어렵다.

본 논문에서는 이러한 한계를 극복하기 위해, 서로 다른 학습 데이터로 사전 학습된 복수의 생성형 모델을 활용하는 멀티 에이전트 챗봇 토론 프레임워크를 제안한다. 해당 프레임워크를 활용한 서비스를 직접 개발하고, 실험을 통해 모델 다양성에 기반한 답변 개선 가능성을 검증한다.

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2025년도 SW중심대학사업의 결과로 수행되었음(2023-0-00042)

3. 멀티 언어 모델을 활용한 멀티 에이전트 챗봇 토론 프레임워크



[그림 1] 전체 프레임워크 동작 구조도

3.1. 에이전트 챗봇

제안하는 프레임워크는 하나의 질문에 대해 두 종류의 에이전트 챗봇이 독립적인 응답을 생성하고, 이들 간의 일치 여부를 비교한 뒤 필요 시 토론 과정을 거쳐 최종 응답을 도출하는 구조로 구성되어 있다 [그림 1]. 프레임워크는 다음과 같은 네 종류의 에이전트 챗봇으로 구성된다.

기본 챗봇(ChatGPT 4o-mini): 웹 브라우저를 통해 사용자와 직접 상호작용하며, 질문에 대한 응답을 생성한다.

논의 챗봇(ChatGPT 4o-mini, Gemini 2.0-Flash): 해당 프레임워크에는 각기 다른 모델의 두 개의 논의 챗봇이 존재한다. 사용자가 질문을 하면, Gemini 2.0-Flash 모델 기반의 논의 챗봇 B 는 사용자 질문에 대해 독립적인 답변을 생성한다.

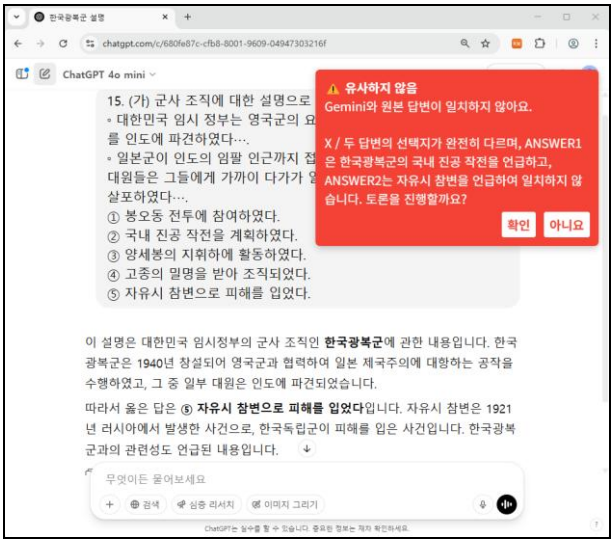
저지 챗봇(Perplexity Sonar): 두 응답 내 정답 선지의 일치 여부를 판단하고, 일치하지 않을 경우 두 논의 챗봇에게 토론을 진행시킨다. 별도의 수치적 유사도 계산 없이 LLM 이 두 응답의 일치 여부를 스스로 판단하도록 설계했다.

모더레이터 챗봇(ChatGPT 3.5-Turbo): 토론이 발생한 경우 토론 내용을 종합하여 최종 응답을 생성한다.

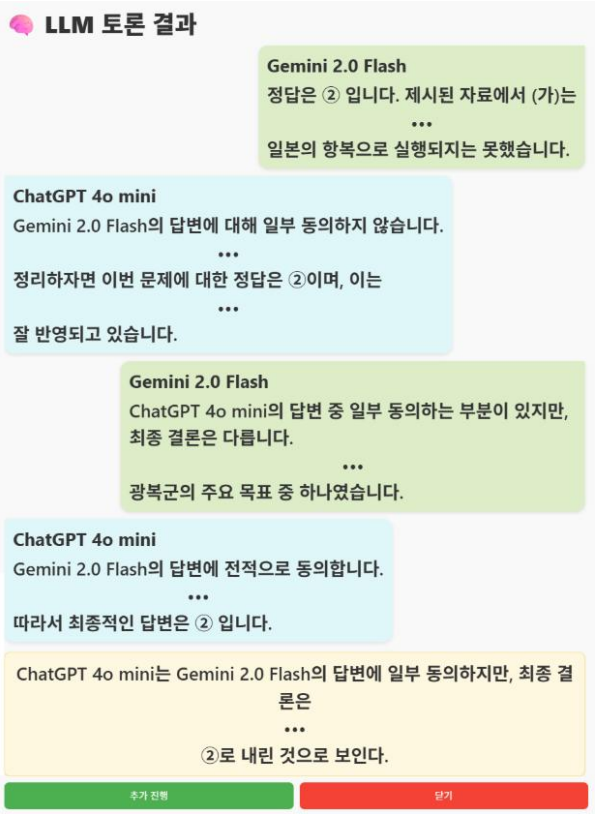
3.2. 프레임워크 동작 과정

사용자가 질문을 제출하면 기본 챗봇이 먼저 답변을 전달하며, 동시에 동일한 질문이 논의 챗봇 B 에도 전달되어 각각 독립적인 답변을 생성한다. 저지 챗봇은 기본 챗봇과 논의 챗봇 B 의 응답을 비교하여 일치 여부를 판단한다. 답변이 일치하면 기본 챗봇의 응답을 최종 채택하고 이후 과정을 종료한다. 만약 두 챗봇의 응답이 일치하지 않을 경우, 기본 챗봇과 동일한 모델의 논의 챗봇 A 와 B 가 번갈아 상대의 답변을 반박하고 스스로 답변을 수정하는 과정을 거친다. 이 한 회의 토론이 끝나면 전체 대화 기록이 모더레이터 챗봇으로 전달되며, 모더레이터 챗봇은 지금까지의 논의를 요약·정리한 후 최종 응답을 생성한다. 최종 응답과 토론 내용이 클라이언트에 전달되며, 사용자가 '추가 진행' 버튼을 클릭하면 동일한 절차를 한 차례 더 수행한다.

4. 환각 탐지 및 답변 개선 시스템 개발



[그림 2] 답변 불일치 시 팝업 창 제시



[그림 3] 논의 챗봇 A, B의 토론 및 종합 답변 생성 예

본 논문에서 제안한 프레임워크를 확장 프로그램으로 구현했다 [1]. 사용자 질문 시, 기본 챗봇과 논의 챗봇 B의 응답이 다르면 우측 상단에 경고 팝업이 뜬다 [그림 2]. 사용자가 '확인' 버튼을 누르면 브라우저 우측 하단에 논의 챗봇A와 B의 토론 내용, 모더레이터 챗봇의 종합 답변이 생성된다 [그림 3].

[1] https://youtu.be/oxzD7FspX6g

5. 멀티 에이전트 챗봇 성능 평가

본 연구가 제안한 프레임워크의 환각 해결 효과 검증을 위해, '2025 학년도 대학수학능력시험' 문항을 질의 데이터셋으로 활용했다. 해당 문항을 논의 챗봇 A(LLM: ChatGPT o4-mini)와 B(LLM: Gemini 2.0-Flash)에 각각 입력해 응답을 수집했다. 그 결과 적어도 한 모델에서 오답이 생성된 66 문항을 선별했고, 이를 본 연구의 실험에 활용했다.

5.1. 단일 모델 기반 멀티 에이전트 챗봇 성능 평가

각기 다른 모델로 구성된 멀티 에이전트 챗봇 토론 프레임워크의 효과를 검증하기 전에, 단일 모델 기반 멀티 에이전트 챗봇에서 환각 문제를 어떻게 해결하는지 확인했다. 본 연구진이 개발한 환각 탐지 및 답변 개선 시스템에서, 논의 챗봇 A와 B를 동일한 ChatGPT 4o-mini 모델로 변경한 후, 수집된 66개 수능 문항에 대해 질의를 수행했다. 이 중 53문항에서 적어도 하나의 챗봇이 오답을 생성했으며 [표 1], 한 모델만 오답을 낸 16문항 중 12문항(75%)에서 토론 후 정답 도출에 성공했다. 반면 서로 다른 오답을 낸 13문항은 토론 후 정답으로 수정되지 않았다(0%).

표 1. 같은 모델 사용 시 변화된 답변 결과

수능 문제 53문항 답변 결과 (개)				
둘 중 한 모델만 틀림		둘 다 틀림		
16		37		
-		다른 오답		같은 오답
		13		24
토론 수행 결과 (개)				
정답	오답	정답	오답	-
12	4	0	13	-

5.2. 멀티 모델 기반 멀티 에이전트 챗봇 성능 평가

서로 다른 모델의 멀티 에이전트 챗봇을 활용했을 때, 총 66문항 중 35문항에서 하나의 모델이 오답을 생성했다 [표 2]. 논의 챗봇 A(ChatGPT 4o-mini)와 B(Gemini 2.0-Flash)의 토론 결과, 이 중 34문항(97%)이 정답을 도출했다. 서로 다른 오답을 낸 28문항 중에도 토론 결과 7문항(25%)이 정답으로 수정됐다. 이는 동일 모델 기반보다 더 높은 성능을 보여준다.

표 2. 다른 모델 사용 시 변화된 답변 결과

수능 문제 66문항 답변 결과 (개)						
둘 중 한 모델만 틀림				둘 다 틀림		
35				31		
ChatGPT만 정답	Gemini만 정답			다른 오답		같은 오답
7	28			28		3
토론 수행 결과 (개)						
정답	오답	정답	오답	정답	오답	-
7	0	27	1	7	21	-

6. 결론

본 연구에서는 서로 다른 데이터로 학습된 멀티 언어 모델을 활용한 멀티 에이전트 챗봇 토론 프레임워크를 제안하고, 이를 기반으로 환각 현상을 탐지·교정하는 시스템을 개발했다. 2025학년도 대학수학능력시험 문항을 대상으로 성능을 검증한 결과, 단일 언어 모델 기반 멀티 에이전트보다 각기 다른 모델 기반의 멀티 에이전트 간 토론에서 답변 개선 효과가 훨씬 우수함을 확인했다. 그러나 두 논의 챗봇이 모두 동일한 오답을 제시할 경우 토론으로도 정답을 도출할 수 없다는 한계가 남아 있다. 이에 향후 연구에서는 다양한 학습 데이터와 모델을 가진 추가 에이전트를 통합하고, 외부 지식베이스를 활용해 모든 챗봇이 동일하게 오답을 제시하는 경우에도 스스로 오류를 검증·교정할 수 있는 메커니즘을 도입함으로써 환각 교정 성능을 향상시킬 계획이다.

참고문헌

[1] John Lincoln. "62% of People Now Use ChatGPT or Google Gemini to Find a Product or Service, Why This Will Change Marketing". Ignite Visibility. 2024.

[2] Farquhar et al. Detecting hallucinations in large language models using semantic entropy. Nature 630.8017. 625-630. 2024.

[3] Sun et al. AI Hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications 11.1 1-14. 2024.

[4] Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in neural information processing systems. 33. 9459-9474. 2020.

[5] Sun et al. Contrastive learning reduces hallucination in conversations. In Proceedings of the AAAI Conference on Artificial Intelligence. 37. 11. 13618-13626. 2023.

[6] Manakul et al. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. 2023.

[7] Song et al. RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. 1548-1558. 2024.

[8] Zhao et al. Medico: Towards Hallucination Detection and Correction with Multi-source Evidence Fusion. 2024.

[9] Sun et al. Towards detecting LLMs hallucination via Markov chain-based multi-agent debate framework. ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1-5. 2025.