

# 2023 (사)ICT플랫폼학회 추계학술대회 자료집

**행사일시** 2023년 12월 7일(목) 14시

**행사장소** 동국대 충무로영상센터 7층

**주최/주관** (사)ICT플랫폼학회  
동국대학교 인공지능협업센터, 강남대 AI.SW융합연구소

**후원** (주)LG CNS, (주)SK브로드밴드, 대신정보통신(주), 세림티에스지(주),  
(주)스페이스앤빈, (주)시아인사이트, 쌍용정보통신(주), 아이씨티웨이(주),  
(주)올포랜드, (주)이액티브, (주)인라스, (주)진인프라, (주)한국IT컨설팅

# 인공지능 할루시네이션에 대응하는 질의개선

<sup>1</sup> 이현정, <sup>2</sup> 홍세영, <sup>3</sup>박찬솔, <sup>4\*</sup>김영철

## Query Improvement against a Artificial Intelligence Hallucination in Large Language Model

<sup>1</sup>Hyun Jeong Lee, <sup>2</sup>Se Young Hong, <sup>3</sup>Chansol Park and <sup>4\*</sup>R. Young Chul Kim

### 요 약

현재 ChatGPT를 비롯한 인공지능 대화 시스템이 급격히 발전되고 있다. 이에 따라 교육, 창작, 검색 등의 다양한 분야에서 인공지능 대화 시스템의 사용이 활발하다. 하지만 대규모 언어 모델(Large Language Model, LLM)은 완벽하지 않다. 특히 LLM의 출력이 정확하지 검증되지 않은 것은 큰 문제이다. 실제로 인공지능 할루시네이션과 같이 실제 패턴이나 객체에 부합하지 않는 무의미하거나 부정확한 결과를 반환하는 경우가 빈번하다. 본 논문에서는 이러한 문제를 해결하기 위해 체계적인 프롬프트 작성 방법을 제안한다. 정확한 질의 프롬프트를 통해 인공지능이 정확한 답변을 할 수 있도록 유도한다. 이를 통해 일반적인 사용자도 쉽게 인공지능 할루시네이션을 회피할 수 있을 것으로 기대한다.

### Key words

*Artificial Intelligence, Large Language Model, ChatGPT, AI Hallucination*

## I. 서 론

본 논문은 2023년 홍익대학교 소프트웨어융합학과 종합 설계 과목에 결과물이다.

최근, LLM의 성장이 가파르다. 그 중에서도 생성형 언어 모델을 이용한 ChatGPT는 교육, 창작, 검색 등 다양한 분야에서 응용되고 있다. 하지만 LLM과 관련된 다양한 종류의 문제점들 또한

존재한다. 이러한 문제점들 중 하나는 생성된 답변의 정확도 문제이다. 그 중에서도 인공지능 할루시네이션 문제는 모델이 실제 패턴이나 객체에 부합하지 않는 무의미하거나 부정확한 결과를 생성하는 현상을 말한다. 본 논문에서는 인공지능 할루시네이션 문제를 해결하기 위해 체계적인 프롬프트 작성 방법을 제안 한다. 이 방법은 생성형 인공 지능의 출력이 사용자의 의도에 더 잘 부합하도록

<sup>1</sup> 홍익대학교 소프트웨어융합학과, 학부생 (hyunjlee@g.hongik.ac.kr)

<sup>2</sup> 홍익대학교 소프트웨어융합학과, 학부생 (C089084@g.hongik.ac.kr)

<sup>3</sup> 홍익대학교 소프트웨어융합학과, 석사과정 (C2193102@g.hongik.ac.kr)

<sup>4</sup> 교신저자 홍익대학교 소프트웨어융합학과, 교수 (bob@hongik.ac.kr)

설계되어, 생성형 인공 지능의 사용성과 효율성을 높인다. 또한 잘못된 프롬프트로 인한 할루시네이션 문제를 해결할 수 있다.

본 논문은 다음과 같이 구성된다. 2장에서는 LLM과 인공지능 대화 시스템 그리고 인공지능 할루시네이션의 문제에 대해 설명한다. 3장에서는 본 논문에서 제안하는 체계적인 프롬프트 작성 방법에 대해 소개한다. 마지막으로, 4장에서는 본 논문의 결론과 향후 연구에 대해 언급한다.

## II. 관련연구

### 2.1 대규모 언어 모델

LLM은 대량의 텍스트 데이터를 학습하여 자연어 이해와 생성 작업을 수행할 수 있다[1]. LLM의 규모는 국내외에서 경쟁적으로 확장되고 있으며, 논문 작성, 프로그래밍, 언어 번역, 교정, 콘텐츠 생성과 같은 다양한 방면에 이용되고 있다.[2].

### 2.2 인공지능 대화 시스템의 현황

인공지능 대화 시스템은 음성 또는 문자를 사용해 사람과 컴퓨터 프로그램이 대화를 통하여 특정 작업을 수행하도록 만들어진 시스템이다. 현대의 인공지능 대화 시스템은 경쟁적으로 확장되며, ChatGPT와 같은 인공지능 대화 시스템은 다양한 분야에서 활용되고 있다. ChatGPT 이외에도 META의 LLAMA2, 구글 BARD와 같은 다양한 모델이 개발 되었다. 인공지능이 생성한 답변은 논리적으로 보일 수 있지만, 인터넷에서 수집된 데이터 셋으로 사용자와 관련된 편향성을 가지고 있을 가능성이 존재한다[2].

### 2.3 인공지능 할루시네이션

ChatGPT는 종종 사람들의 질문에 정확한

답을 알지 못하면서도 정확한 답을 찾는 것처럼 답변하는 경우가 있다. 이를 할루시네이션 현상이라고 한다. 대화형 인공지능의 할루시네이션은 편향되거나 부정확한 데이터, 입력에 대한 정확한 이해가 아닌 통계적 패턴을 기반으로 응답을 생성하는 모델의 경향으로 인해 발생 가능하다[3]. 다시 말해 훈련데이터의 문제, 확률적 추론의 문제, 진실 데이터의 부재, 모델의 복잡성 등 다양한 원인으로 인해 부정확한 답변을 제시하게 된다.

순다르 피차이 구글 최고경영자 (CEO)는 "바드 등 AI 챗봇에서 발생하는 할루시네이션 원인과 해결책을 그 누구도 제시하지 못하고 있다"며 "앞으로 AI 개발자 최우선 임무가 될 것"이라고 주장했다[4].

표 1은 LLM 할루시네이션 오류 사례이다. 현재로서 GPT-4가 인간과 비슷한 판단력을 지닌 범용 인공지능 (AGI) 수준에 도달하지 못하고 있다. 특히 문장의 이해 능력이 제한적이며, 정보가 부족하거나 사람들이 많이 물어보지 않은 질문에는 적절한 답을 생성하지 못하는 제한된 성능을 보인다.

표 1. LLM 할루시네이션 오류 사례

질문	<p>공쥐팔쥐 이야기 알려줘</p>
ChatGpt 대답	<p>"공쥐팔쥐"는 한국전래동화중하나로, 두마리의작은동물인공쥐와팔쥐가등장하는이야기입니다.이이야기는어린이들을대상으로한교훈적인내용을담고있습니다.아래는"공쥐팔쥐"이야기의일반적인버전입니다.</p> <p>옛날옛적에공쥐와팔쥐라는두마리의친구가살고있었습니다.어느날,둘은함께밭에나가농사를짓기로결심했습니다.공쥐는콩을심고,팔쥐는팥을심었습니다.</p> <p>시간이흘러가을이되었을때,공쥐의밭에는푸른콩이아름답게자라고,팔쥐의밭에는붉은팥이풍성하게열렸습니다.두마리는자신의농사결과에자랑스러워하며서로에게자랑을했습니다.</p> <p>그러나어느날,공쥐는팔쥐를자신의집에초대했습니다.공쥐는팔쥐에게콩으로만든다양한음식을대접했습니다.그리고팔쥐는공쥐를자신의집에초대하여팥으로만든다양한음식을대접했습니다.</p> <p>두마리는서로의집에서즐거운시간을보내며농사의결과물을나누었습니다.이를통해둘은서로의노력과공헌을인정하고협력의중요성을깨달았습니다.</p> <p>"공쥐팔쥐"는이런협력과공유의가치를가르치며,서로다른노력에도모성성한결과를가져오는소중함을전하는이야기입니다.</p>

표 2. 관련 배경정보를 통한 질의개선 예시

질문`	<p>공쥐팔쥐는 공쥐와 팔쥐라는 두 자매의 이야기로 계모에 대한 모순된 대우와 학대로 자매간의 차별에 대한 비판적 시각을 제시하는 이야기야 계모 밑에서 갖은 학대를 받던 공쥐가 고귀한 인물과 혼인하게 되고, 공쥐를 괴롭히던 팔쥐와 계모는 처벌받는다는 내용의 설화이지 다시한번 공쥐팔쥐 이야기를 알려줄래?</p>
ChatGpt 대답	<p>공쥐와 팔쥐가 자매인 버전의 "공쥐팔쥐" 이야기는 한국의 전래동화 중 하나로, 주로 공쥐의 어머니와 계모에 대한 모순된 대우, 그리고 자매간의 협동과 도움을 강조하는 내용을 담고 있습니다. 이 이야기는 전통적인 가치관과 도덕적인 교훈을 전하는데 중점을 두고 있습니다.</p> <p>이야기의 주요 흐름은 다음과 같습니다: 옛날 옛적에 공쥐와 팔쥐는 어머니와 함께 살고 있었습니다. 그러나 어머니는 공쥐를 편애하며 팔쥐에게는 불공평한 대우를 했습니다. 어머니는 공쥐에게는 좋은 음식과 편안한 공간을 제공하면서도 팔쥐에게는 나쁜 대접을 했습니다.</p> <p>한 날, 공쥐는 산에서 우연히 선녀를 만나게 됩니다. 선녀는 공쥐에게 도움을 주고 감사의 표시로 보석들을 선물했습니다. 공쥐는 선녀에게 감사의 인사를 전하고, 선녀는 "어떤 어려움에 처했을 때, 내 이름을 불러 도움을 청하면 나타날 것이다" 라고 말합니다.</p> <p>이후, 공쥐는 어머니의 대우에 불만족하며 선녀를 불러 도움을 청합니다. 선녀는 공쥐를 도와 어머니에 대한 불공평한 대우를 바로잡아줍니다. 이에 어머니는 공쥐와 팔쥐에게 공평한 대우를 하게 되지만, 어머니의 태도는 변하지 않습니다.</p> <p>이후, 공쥐는 팔쥐를 돕기 위해 선녀를 다시 부르고, 선녀의 도움 으로 팔쥐에게도 공평한 대우가 이루어집니다. 공쥐와 팔쥐는 서로 협동하여 어머니와 함께 행복한 삶을 살게 되며, 결국 공쥐는 선녀와의 인연을 통해 혼인하게 됩니다.</p> <p>이 이야기는 가치관의 변화와 협동의 중요성, 그리고 선행과 감사의 가치를 강조하는 전래동화로, 어머니와 계모에 대한 불공평한 대우를 극복하고 가족간의 협력을 통해 행복한 결말을 이루어냅니다.</p>

### Ⅲ. 할루시네이션 해결을 위한 질의개선 연구

할루시네이션 오류를 해결하기 위해 모델을 개선하거나, 대화 시스템을 수정 할 수 있다. 하지만 일반적인 사용자가 적용하기에는 어려운 방법이다.

일반적인 사용자가 인공지능의 할루시네이션을 회피하기 위한 방법으로 질의 프롬프트 개선을 제안한다. 질의 프롬프트 개선을 통해 인공지능 대화 시스템으로부터 정확한 답변을 유도할 수 있다.

### 3.1 관련 배경정보 추가

표 2는 관련 핵심 키워드를 프롬프트에 포함 하여 ChatGPT 답변의 정확도를 향상한 것이다. 배경정보를 프롬프트에 포함하면 모델이 대화의 맥락과 의도를 더 잘 이해할 수 있게 되어 모델은 더 일관된 답변을 생성하고, 사용자의 의도에 더 잘 부합하는 내용을 생성할 가능성이 높아진다. 예를 들어, "한국 전래동화 중에서 콩쥐팥쥐에 대해 알려줘"라는 프롬프트를 사용하면, 모델은 사용자가 전래동화에 대한 정보를 요청하는 것임을 명확히 이해하고 이에 맞는 정보를 생성한다.

### 3.2 관련 핵심 키워드 추가

표 3. 핵심 키워드를 통한 질의개선 예시

질문	콩쥐팥쥐는 콩쥐 와 팥쥐라는 사람의 이야기로 자매간의 불공평한 대우에 관한 전래동화야 다시한번 콩쥐팥쥐 이야기에 대해서 알려줘
ChatGpt 대답	"콩쥐팥쥐"는 한국의 전래동화 중 하나로, 자매 간의 불공평한 대우에 대한 이야기를 다룹니다. 이 이야기는 한국의 전통적인 가치관과 더불어 도덕적인 교훈을 담고 있습니다. 이 동화는 주로 두 자매, 콩쥐와 팥쥐에게 중심을 둡니다. 두 자매는 모두 열심히 노력하며 살아가지만, 환경이나 상황에 따라서 각자의 운명이 다르게 전개됩니다. 대개 콩쥐는 불운하게, 팥쥐는 행운을 맞이하는 경우가 많습니다. 이 동화는 주로 불행한 처지에서 희망을 잃지 않고 노력하는 데 중점을 두며, 어떤 상황에서도 긍정적으로 행동하고 인내심을 가지는 것의 중요성을 강조합니다. 또한, 자매 간의 차별에 대한 비판적 시각을 제시하여 공평하고 인간다운 사회를 지향하는 메시지를 전합니다.

표 3은 관련 핵심 키워드를 프롬프트에 포함하여 ChatGPT 답변의 정확도를 향상한 것이다. 프롬프트에 핵심 키워드를

포함시키면 ChatGPT는 사용자의 의도와 관련된 핵심 주제를 명확히 이해하고, 그에 따른 응답을 생성하는 데 도움이 된다. 특히 핵심 키워드는 모델이 어떤 정보를 포함해야 하는지에 대한 신호를 제공하며, 이를 통해 모델이 더 명확하고 정확한 답변을 생성할 수 있다. 예를 들어, "콩쥐팥쥐"가 사람인 것에 관한 정보를 얻고 싶다면, "콩쥐와 팥쥐라는 사람의 이야기로 자매 간의 협동과 도움을 강조한 전래동화"라는 키워드를 프롬프트에 포함시킬 수 있다. 이렇게 하면 모델은 주어진 키워드를 중심으로 정보를 제공하려고 시도하게 된다.

ChatGPT를 사용할 때 가장 중요한 것은 프롬프트 작성이다. 관련 배경 정보를 제공하면 ChatGPT가 질문의 맥락을 더 잘 이해 가능하다. 또한, 내가 원하는 바를 명확하게 질문해야 정확한 답변을 받을 수 있다.

## IV. 결 론

본 논문에서는 인공지능 할루시네이션 문제를 해결하기 위해 프롬프트 질의 개선 방법을 제안했다. 프롬프트를 명확하게 주는 간단한 방법을 통해 인공지능 대화 시스템이 정확한 답변을 할 수 있도록 유도할 수 있다. 이를 통해 일반적인 사용자도 인공지능 대화 시스템의 할루시네이션 문제를 간단하게 회피할 수 있을 것으로 기대한다. 추후 연구 과제로는 할루시네이션 패턴을 분석하여 이를 해결할 수 있는 질의 프롬프트 구조를 설계할 예정이다.

## 참 고 문 헌

[1] 이세훈, 최정빈, 백영태, 윤선호. (2023.07).