# 초거대 언어모델의 재치에 관한 고찰: 수수께끼 해결 능력을 중심으로

어수경[1○], 박찬준[2], 문현석[1], 서재형[1], 허윤아[1,3*], 임희석[1,3*]
[1]고려대학교 컴퓨터학과, [2]업스테이지 [3]Human-inspired AI 연구소
{djtnrud,glee889,seojae777,yj72722,limhseok}@korea.ac.kr
bcj1210@naver.com

# A Study on Proficiency in Solving Riddles of Large Language Model

Sugyeong Eo[1○], Chanjun Park[2], Hyeonseok Moon[1], Jaehyung Seo[1], Yuna Hur[1,3*], Heuiseok Lim[1,3*]
[1]Department of Computer Science and Engineering, Korea University, [2]Upstage [3]Human-inspired AI Research

**요 약**

초거대 언어모델은 과연 수수께끼 문제에 재치있는 답변을 할 수 있을까? 최근 초거대 언어모델(Large language model, LLM)은 강력한 성능 및 유저 만족도를 보이며 세계의 이목을 집중시키고 있다. 여러 태스크들에 대한 정량 평가를 넘어서 최근에는 LLM의 창의력 및 고도화된 언어능력을 평가하는 연구들이 등장하고 있다. 본 논문에서는 이러한 연구 흐름에 따라 LLM의 재치에 관해 고찰해본다. 이때 재치를 평가하기 위한 태스크로 이를 요구하는 말놀이인 수수께끼를 활용한다. 본 논문은 LLM이 수수께끼를 잘 수행하는지를 모델 추론을 통해 평가하며, 모델 추론 시 활용되는 프롬프트들의 성격에 따른 성능 변화를 관찰한다. 또한 수수께끼의 종류에 따른 모델의 능력을 비교 분석하며 LLM의 추론 결과에 대한 오류 분석을 수행한다. 본 논문은 실험을 통해 GPT-4가 가장 높은 성능을 보이며, 설명글이나 데이터 예시를 추가할 시 성능을 한층 더 향상시킬 수 있음을 확인한다. 또한 단어 기반보다는 특성 기반의 수수께끼에 더욱 강력한 성능을 보이며, 오류 유형 분석을 통해 LLM이 환각(hallucination) 문제와 창의력을 동시에 가지고 있다고 분석한다.

주제어: 자연어처리, 초거대 언어모델, 언어모델, LLM Creativity

## 1. 서론

최근 초거대 언어모델 (Large language model, LLM)은 다양한 태스크들을 처리할 뿐만 아니라 맥락내 학습(in-context learning)을 통해 사전학습 시 보지 못했던 새로운 태스크들에 대한 수행이 가능해졌다. 이에 따라 LLM의 창의력 및 고도화된 언어능력을 테스트하는 연구들도 등장하고 있다. 이들은 LLM이 단순 피상적인 언어의 이해 및 생성을 넘어서 의미 및 맥락을 이해하고 이에따라 창의적이거나 유창한 답변을 얼마나 잘 생성하는지를 평가한다 [1, 2, 3, 4].

본 논문에서는 이러한 연구 흐름에 탑승하여 LLM의 재치에 관해 고찰해본다. 재치는 추상적 사고 능력 및 창의력을 바탕으로 하는 능란한 솜씨나 말씨로 간주되며, 본 논문은 LLM이 이러한 능력을 지니고 있는지를 평가한다. 본 논문은 지능적인 유머로써 재치가 단어 놀이와도 연결된다는 점에서[1] 한국의 대표적인 말놀이중 하나인 수수께끼를 평가 대상으로 선정한다. 수수께끼의 경우 재치있고 유머러스한 질의-응답 쌍으로 구성되며, 질의에 대한 모델 추론 결과를 통해 LLM의 창의력이나 추상적 사고 능력을 판단할 수 있다. 더 나아가 수수께끼의 정답을 도출해내는 과정에서 추론 능력 및 메타언어적 능력(meta-linguistic ability)을 함께 요구하기 때문에 한층 더 고차원적인 언어 능력을 평가할 수 있게 된다.

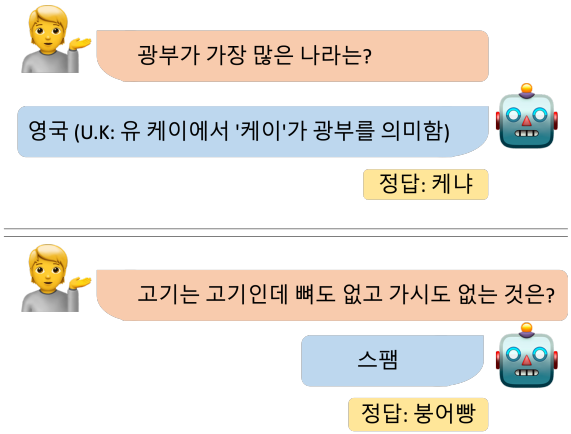


그림 1. 수수께끼에 문제에 대한 LLM 추론 예시

본 논문의 실험은 프롬프트의 형태로 LLM에게 수수께끼 질문을 제시하고 LLM이 추론을 통해 결과를 반환하는 방식으로 진행된다. 이 과정에서 총 세 가지 논점에 대해 정량적으로 평가하고자 한다: (1) LLM의 수수께끼 해결 능력은 어느 정도이며, 어떤 LLM이 가장 우수한 성능을 보이는가? (2) 데이터 예시 제공이 모델 추론에 어떤 영향을 주는가? (3) 수수께끼 설명글의 포함이 모델 추론에 어떤 영향을 주는가?

첫 번째 논점으로부터 LLM의 수수께끼 해결 능력은 과연 어느 정도인지를 알아보고자 하며, 어떤 LLM이 가장 높은 성능을 냈는지에 대해 알아본다. 비교를 위한 모델 선정으로는

*교신저자(Corresponding author)
[1]https://en.wikipedia.org/wiki/Wit

대표적인 생성모델인 InstructGPT [5], ChatGPT [6], GPT-4 [7]를 활용한다. 두 번째와 세 번째 논점은 프롬프트 엔지니어링(prompt engineering)의 관점에서 예시 데이터를 포함하는 것 또는 설명글(description)을 포함하는 것이 각각 모델의 추론 성능에 어떤 영향을 주는지를 파악하고자 한다. 실험 결과에 대한 분석에서는 본 논문은 수수께끼를 크게 단어 기반 및 특성 기반의 두 종류로 분류하여, 어떤 분류에서 더 높은 오류율을 보이는지를 확인한다. 마지막으로 LLM이 수수께끼에 대한 오답을 제시했을 때 어떤 오류들을 생성하는지 오류 유형을 분석해보고자 한다.

실험 결과 정량 평가에서 GPT-4가 일관적으로 가장 우수한 성능을 기록했으며, 전반적으로 데이터의 예시를 늘렸을 때 전반적으로 성능 향상과 양의 상관관계를 보였다. 또한 설명글을 추가하는 것 역시 성능 향상을 가능하게 했다. 분석에서는 LLM이 특성 기반의 문제를 더욱 잘 수행한 것을 확인하며, 단어 기반의 문제를 더욱 어려워하는 경향을 확인한다. 분석을 통해 발견한 특징적인 결과는 LLM이 상당한 창의력을 지니나 여전히 환각과 같은 문제에서 벗어나지 못하는 점이었다. 이는 그림 1에서도 잘 드러난다. LLM의 높은 재치는 창의력을 요구하는 태스크나 창작의 관점에서 휴먼과 AI의 합작(Human-AI collaboration)을 통한 상당한 이점을 볼 수 있다. 이에 본 연구를 바탕으로 재치 및 창의력 등 생각의 발산으로서의 LLM에 대한 연구가 지속되기를 바라는 바이다.

## 2. 관련 연구

**초거대 언어모델** 최근 초거대 언어모델 (Large language model, LLM)은 우수한 성능 향상과 사용자 만족도를 기록하며 세계의 이목을 집중시키고 있다. LLM은 수십에서 수천억 개의 파라미터를 지닌 초거대 사이즈의 언어모델로, 사전학습 언어모델 (Pre-trained language model)과 비교하여 정량, 정성 및 휴먼 등 전반적인 평가에서 우수한 성능을 기록했다. 특히나 T5 [8]를 시작으로 GPT-3 [9], OPT [10], FLAN-T5 [11], LLaMA [12, 13], InstrucGPT [5], ChatGPT [6], GPT-4 [7] 등 다양한 LLM들이 연이어 등장하면서 질의응답, 대화시스템 및 요약 등의 태스크에서 그 기록들을 경신하고 있다.

최근에는 LLM의 창의력 및 고도화된 언어능력을 테스트하는 연구들도 등장하고 있다. 예를 들어 LLM의 시나 스토리 등 문학과 관련한 창작 능력을 평가한다 [1, 2, 3]. 맥락에 따라 말의 의도를 파악하는 화용적 표현들을 얼마나 잘 이해하였는지에 대해 연구하기도 한다 [4, 14]. 본 논문에서는 LLM의 재치에 관해 연구하며, 이를 얼마나 지닌 채로 수수께끼를 잘 풀어내는지에 대해 알아보고자 한다.

**수수께끼** 수수께끼는 불가능한 단어 조합 및 비유적인 표현들을 바탕으로 이들이 가리키는 정답을 유추해내는 말놀이이다 [15, 16]. 따라서 일반적인 사고의 흐름을 깨는 질문 및 응답으로 구성되어 있기에 말풀이 과정에서 추상적 사고 능력이나 창의력이 기반되는 재치가 요구된다. 특히나 수수께끼의 경우 추가적으로 정답 도출 과정에서 추론, 문화 및 언어적 특성에 대해 주목하는 메타언어적 능력 등도 함께 요구한다 [17, 18, 19]. 이에 따라 수수께끼를 본 실험에 활용한다면 한층 더 고도화된 언어 능력의 관점에서 LLM을 평가할 수 있는 척도가 될 것으로 판단한다.

본 논문은 수수께끼를 크게 (1) 단어 기반, (2) 특성 기반의 두 종류로 나눈다. 먼저 전자의 경우 “Q: 왕이 넘어지면 어떻게 될까? A: 킹콩” 처럼 단어나 단어 내 일부 형태소를 활용한 언어유희를 의미한다. 후자의 경우 “Q: 거꾸로 들고 보나 바로 들고 보나 바로 보이는 것은?” A: 거울” 처럼 단어의 특성을 기반으로 하는 퀴즈를 의미한다.

## 3. LLM의 수수께끼 해결 능력

### 3.1 실험 세팅

**데이터** 수수께끼 데이터는 웹 크롤링을 통해 수집하였다. 수수께끼 데이터의 수집 이후, 분석을 위해 이를 두 가지 종류인 단어 기반, 특성 기반으로 분류하는 작업을 수행한다. 이에 따라 말 기반 679, 특성 기반 130개로, 총 809개의 수수께끼 데이터를 생성한다.

**모델** 실험에 활용한 LLM의 경우 InstructGPT, ChatGPT, GPT-4를 활용하였다. 위 모델들의 경우 모두 OpenAI에서 공개한 모델이며, InstructGPT의 경우 휴먼 피드백(human feedback)을 활용하여 강화학습 방식의 인스트럭션 튜닝(instruction tuning)을 수행한다. ChatGPT 모델은 대화를 가능하게 하도록 적응(adapt)시킨 모델이며, 사람과의 대화에서 놀라운 능력을 보였다. GPT-4는 이미지와 텍스트 모두를 입력받을 수 있는 멀티모달(multi-model)로서 다양한 태스크에서 휴먼 레벨의 성능을 보였다. 본 논문에서 활용한 모델의 버전은 InstructGPT (text-davinci-003), ChatpGPT (gpt-3.5-turbo), GPT-4(gpt-4) 이며, OpenAI에서 제공한 API를 활용하였다.

**평가지표** 평가지표로는 질의응답(question-answering, QA)에서 활용되는 대표 지표인 exact-match (EM)와 F1-score를 활용한다. EM의 경우 정답과 모델 추론 결과의 정확한 단어 일치도를 평가한다. F1-score의 경우 정답을 기준으로 모델 추론 결과와의 단어 일치도를 계산하는 재현율(recall)과 모델 추론 결과를 기준으로 정답과의 단어 일치도를 계산하는 정확률(precision) 간의 조화 평균을 계산한 값이다.

### 3.2 수수께끼 특화 프롬프트 엔지니어링

입력되는 프롬프트는 설계하는 방법에 따라 LLM의 성능은 상당한 편차를 보인다. 따라서 프롬프트 엔지니어링 과정에서 데이터 예시 추가 및 수수께끼 설명글 추가의 총 두 가지 정보를 추가한다.

**데이터 예시** 먼저 프롬프트의 입력 요소 중 하나로 데이터의 예시를 선택한다. 즉 LLM이 수수께끼 질의-응답을 직접적으로 학습하지 않았기 때문에 이와 유사한 예시들을 프롬프트에 추가해주는 방법을 활용한다. 이때 데이터 예시의 개수에 따라 제로샷 (zero-shot) 또는 퓨샷 (few-shot)으로 구분할 수 있는데, 샷을 점진적으로 늘려가면서 데이터 예시의 개수가 모델 추론 성능에 어떤 영향을 미치는지 확인한다. 프롬프트에 추가할 데이터 예시 선정 시에는 수집된 809개의 데이터 예시와 겹치지 않도록 새로운 수수께끼 질의-응답 쌍을 수집한다. 이때 수수께끼의 각 종류별로 8개씩 총 16개의 예시 데이터를 수집한다. 최종적으로 생성한 프롬프트 중, 데이터 예시로 구성된 2-shot 프롬프트 예시는 다음과 같다: *"다음 수수께끼 질문에 대한 정답은?: Q: 해를 취재하는 사람은? A: 해리포터, Q: 눈은 3개 다리는 1개인 것은? A: 신호등, Q:"*

**수수께끼 설명글** 프롬프트의 입력 요소 중 하나로 수수께끼에 대한 설명글을 입력으로 넣어주는 방법을 선택한다. 이는 수수께끼에 대한 정의를 통해 LLM이 어떤 태스크인지를 판단할 수 있도록 유도하고자 함이다. 다음은 수수께끼의 설명글 전문이다: *"수수께끼는 질문에 대해서 재치 있는 대답을 요구하는 말장난을 이용한 퀴즈다. 다만 보통 퀴즈와는 달리 정답이 사실에 근거한 것보다, 말의 의미를 억지로 가져다 붙이거나 동음이의어를 이용한 익살이나 농담인 경우가 많다. 음운을 되풀이하거나, 무언가에 빗대어 표현한다. 돌려서, 단어에 의해 완곡적으로 알 수 있게 하는 것도 수수께끼라고 한다. 다음 수수께끼 질문에 대한 정답은? Q: "*

## 4. 실험 결과

실험에서는 서론에서 제시한 세 가지 논점을 활용하여 실험 결과에 대해 결과를 정량 평가한다. LLM의 수수께끼 해결 능력에 대한 실험 결과는 표 1과 같다.

### 4.1 어떤 LLM이 가장 높은 성능을 보이는가?

본 논문에서 활용한 InstructGPT, ChatGPT, GPT-4중 모든 세팅에서 GPT-4가 가장 우수한 성능을 보였다. 0점대와 1에서 3점대를 보이는 InstructGPT 및 ChatGPT와 비교하여, 최대 EM 6.304 및 F1 6.919의 성능을 보이며 확연한 성능 차이가 나타났다. GPT-4 모델의 경우 창의력을 요구하는 태스크가 아닌 QA, 대화 및 요약 등의 일반적인 태스크에서도 가장 높은 성능을 보이는 모델인데, 이와 일관된 결과로 수수께끼를 활용한 실험에서도 가장 높은 성능을 보였다.

| Shot | Model | w/o Desc | | w/ Desc | |
|---|---|---|---|---|---|
| | | EM | F1 | EM | F1 |
| 0 | InstructGPT | 0.124 | 0.532 | 0.247 | 0.507 |
| | ChatGPT | 0.007 | 1.323 | 1.112 | 1.497 |
| | GPT-4 | 3.461 | 4.045 | 4.203 | 4.887 |
| 1 | InstructGPT | 0.124 | 0.433 | 0.000 | 0.297 |
| | ChatGPT | 1.854 | 2.460 | 1.483 | 2.011 |
| | GPT-4 | 5.068 | 5.715 | 5.810 | 6.700 |
| 2 | InstructGPT | 0.865 | 1.133 | 0.371 | 0.453 |
| | ChatGPT | 2.472 | 2.839 | 1.978 | 2.525 |
| | GPT-4 | 5.068 | 5.787 | 5.686 | 6.336 |
| 4 | InstructGPT | 0.618 | 0.977 | 0.494 | 0.715 |
| | ChatGPT | 0.026 | 3.390 | 2.596 | 3.199 |
| | GPT-4 | **5.192** | **5.870** | 4.326 | 5.297 |
| 8 | InstructGPT | 0.618 | 0.853 | 0.004 | 0.519 |
| | ChatGPT | 2.225 | 2.510 | 1.854 | 2.304 |
| | GPT-4 | 4.944 | 5.455 | 6.180 | 6.861 |
| 16 | InstructGPT | 0.618 | 0.742 | 0.618 | 0.618 |
| | ChatGPT | 1.978 | 2.474 | 1.854 | 2.277 |
| | GPT-4 | 4.944 | 5.500 | **6.304** | **6.919** |

표 1. LLM의 수수께끼 문제 해결 능력 실험 결과표

그러나 전반적으로 LLM의 수수께끼 해결 능력 성능은 매우 저조한 점수대로 분포되어 있다. 가장 우수한 성능을 내는 GPT-4 역시 EM 및 F1 스코어에서 10% 미만의 낮은 결과를 보였다. 이를 통해 LLM은 수수께끼와 같은 재치를 요구하는 태스크는 잘 수행하지 못하는 경향을 보임을 확인한다. 또한 창의력이나 추상적 사고 능력 역시 한국어에서는 전반적으로 저조한 것을 확인할 수 있었다.

### 4.2 데이터 예시 제공이 성능에 영향을 줄 수 있는가?

데이터 예시의 제공은 모델로 하여금 태스크가 어떤 식으로 진행되는지를 알 수 있는 우수한 참고자료가 될 수 있다. 이러한 데이터의 샷을 늘렸을 때는 LLM의 성능에도 어떠한 영향을 주는지를 확인해본다. 실험 결과 샷을 늘렸을 때 GPT-4를 기준으로 전반적으로 성능이 향상되었다. 0-shot과 비교하여 설명글을 추가했을때나 추가하지 않았을 때 모두 각각 EM 기준 3.461에서 최대 4-shot 5.192로, 4.203에서 최대 16-샛 6.304로 성능이 오르는 것을 확인할 수 있었다. InstructGPT 및 ChatGPT에서

| Model | Δ EM | Δ F1 |
|---|---|---|
| InstructGPT | -0.205 | -0.260 |
| ChatGPT | +0.386 | -0.197 |
| GPT-4 | +0.639 | +0.771 |

표 2. 수수께끼 설명글 추가에 따른 성능 변화 비교. 셀 중 푸른색은 성능의 감소를, 붉은색은 성능의 증가를 나타냄

도 제로샷 성능과 퓨샷 성능을 비교했을 때 전반적으로 성능이 향상했다. 이를 통해 데이터의 예시를 제공하는 것은 LLM의 추론 능력 향상에 도움을 주는 것을 확인한다. 그러나 샷을 점진적으로 늘렸을 때 성능이 일관적으로 오르는 것은 아니므로 데이터 예시의 수와 모델 추론 성능이 반드시 양의 상관관계를 지니고 있는 것은 아님을 확인하였다.

### 4.3 설명글의 포함이 성능에 영향을 줄 수 있는가?

수수께끼 설명글을 포함하는 것은 수수께끼가 어떤 태스크인지를 LLM으로 하여금 알 수 있게 한다. 이러한 설명글의 포함이 모델의 수수께끼 해결 능력에도 긍정적인 영향을 주는지에 대해 확인해보고자 한다. 수수께끼 설명글을 포함했을 때의 성능 변화는 표 2에 나타나있다. 해당 표는 수수께끼 설명글을 추가했을 때의 성능 변화량에 대한 모델별 평균을 측정한 결과이다. 실험 결과 InstructGPT의 경우 오히려 설명글을 추가하는 것이 성능 감소로 이어졌으며, ChatGPT에서는 EM의 성능에서만 성능 향상이 있었다. 그러나 GPT-4의 경우 설명글을 추가했을 때 EM과 F1 모두에서 각각 +0.639, +0.771로 전반적으로 성능이 향상한 것을 확인할 수 있었다. 특히나 주목할 만한 점은 GPT-4의 경우 설명글의 추가와 함께 16shot을 주었을 때 성능이 가장 많이 향상되는 것을 확인할 수 있었다. 이를 통해 맥락내 학습(in-context learning)이 설명글 추가 및 다양한 데이터 예시 제공을 했을 때 가장 우수하게 가능했음을 확인한다.

## 5. 추가 분석

본 논문은 세 가지 논점에 대해 LLM의 수수께끼 성능과 성능 증감에 영향을 미치는 요소들에 대한 조사를 수행했다. 다음으로는 실험 결과에 대해 두 가지의 분석을 수행한다. 이에 LLM이 수수께끼 중 어떤 종류에 특히나 취약하며, 어떤 오류들을 범하는지를 확인해보고자 한다.

### 5.1 수수께끼의 특성에 따른 성능 비교

수수께끼는 그 특징에 따라 단어 및 특성의 두 분류로 나뉘어진다. 본 논문은 수수께끼의 특성에 따라 데이터를 분류하고, 특성별 EM 및 F1 스코어를 측정한다. 실험에 활용한 모델은 표 1에서 가장 우수한 성능을 보인 GPT-4를 활용하였다.

| 수수께끼의 종류 | EM | F1 | # Data |
|---|---|---|---|
| 단어 기반 | 7.511 | 8.243 | 679 |
| 특성 기반 | 33.846 | 37.671 | 130 |

표 3. 수수께끼의 종류에 따른 GPT-4 모델 성능 비교

수수께끼의 특성에 따른 GPT-4 모델 성능을 비교한 결과는 표 3와 같다. 결과는 상당히 놀라운데, GPT-4 모델은 주로 특성 기반의 수수께끼가 아닌 단어 기반의 수수께끼를 훨씬 더 잘하지 못하는 경향을 확인할 수 있었다. 즉, 단어 기반 수수께끼의 경우 GPT-4는 EM 약 7.5의 성능을, 특성 기반 수수께끼의 경우 약 33.8의 성능을 보였다. 이에 대한 분석으로, 영어에서는 주로 수수께끼가 Riddle로서 특성 기반의 질의-응답과 유사하다는 점을 들 수 있다 [16, 20]. GPT-4는 영어를 주요 언어로 하여 가장 많은 코퍼스를 활용해 학습했으며, 영어권에서는 주로 수수께끼로 특성 기반의 문제를 다루기 때문에 이러한 특성 기반의 수수께끼가 한층 더 학습 시 노출이 많이 되었을 가능성이 있을 것이다. 이에 대한 결과로 단어 기반 및 특성 기반 실험에서 확연한 성능 변화가 나타났다고 분석할 수 있다.

### 5.2 LLM의 오류 유형 분석

마지막으로 본 논문에서는 LLM의 오류 유형에 대해 분석해본다. LLM은 주로 저조한 성능을 보였는데 어떤 유형의 오류를 생성해냈는지에 대하여 질적 분석을 수행하였다. 오류 유형에 대한 분석 결과는 표 4와 같다.

첫 번째 행에서 LLM은 존재하는 엔티티를 추론 결과로 출력했으나, 이는 질의와 전혀 관련없는 결과로 나타났다. 반대로 두 번째 행에서는 질의와는 관련이 있어 보이는 '펭탕'이라는 결과를 출력했지만 이는 존재하지 않는 엔티티로 환각을 일으킨 것을 확인할 수 있다. 질의와 전혀 관련이 없으면서 동시에 존재하지 않는 엔티티를 생성한 결과도 있는데, 서울에서 가장 땅값이 싼 동네를 질문한 질의와 전혀 관련이 없는 '강추( carrot)때문에' 라는 응답을 하였다. 이를 통해 LLM은 내재적인 지식을 활용하여 문제를 풀 수 없는 경우 환각을 다수 생성하는 것을 확인할 수 있었다. 환각은 LLM의 고질적인 문제이며, 창의력이나 재치 등과 같은 발산하는 생각의 영역이 지나친 발산으로 이어질 경우 이러한 문제들이 발생할 수 있다고 분석한다. 따라서 일관성 및 유창성, 연관성을 해치지 않는 선에서의 적절한 재채를 지니는 응답들을 생성해야 할 필요가 있다고 결론짓는다.

네 번째 행은 수수께끼에 대한 태스크 이해를 기반하지 않고 일반적인 문제 해결으로 간주하였을 때의 오답이다. 즉, LLM

| Question | Answer | LLM 추론 결과 | 질의와의 연관성 | 엔티티 존재여부 | 비고 |
|---|---|---|---|---|---|
| 개 가운데 가장 큰 개는? | 안개 | 기린 | ✗ | ✓ | |
| 펭귄 한 마리를 넣고 끓인 탕은? | 설렁탕 | 펭탕 | ✓ | ✗ | 환각 |
| 서울에서 땅값이 제일 싼 동네는? | 일원동 | 강추( carrot) 때문에. | ✗ | ✗ | 환각 |
| 세종대왕이 나온 고등학교는? | 가갸거겨고교 | 없음 (세종대왕의 시대에는 고등학교라는 개념이 없었음) | ✓ | ✓ | 상식적인 답 |
| 고양이를 무서워하지 않는 쥐는? | 박쥐 | 컴퓨터 마우스 | ✓ | ✓ | 재치있는 응답 |

표 4. LLM의 수수께끼 추론 결과에 대한 오류 유형 분석

은 재치있는 정답으로 도출해낸 것이 아닌 상식적인 답을 생성함으로써 오답으로 간주되었다. 마지막으로 주목할만한 점은, 특성 기반의 문제에 대한 LLM의 응답이다. 다섯 번째 행에서 비록 EM, F1 스코어를 통해서는 오답으로 간주되었으나, 질적 분석을 통해 LLM이 상당히 우수한 재치 및 추론 능력을 지니고 있음을 확인할 수 있었다. 이뿐 아니라 "Q: 눈 오는 날만 일하는 사람은? A: 눈사람(정답: 안과의사), Q: 타는 것인데 앞뒤로 못 움직이는 것은? A: 양초(정답: 엘리베이터), Q: 동물원에서 가장 비싼 동물은? A: 금붕어(정답: 백조)" 등 상당히 창의적이고 재치있는 발상을 통한 응답을 생성해냈다. 이에 대해, 표 3에서의 결과처럼 영어권에서는 특성 기반의 수수께끼가 더욱 일반적이며 관련 데이터들을 학습했을 가능성이 있다는 점도 고려할 부분이다. 그러나 이는 확인될 수 없고 특성 기반의 수수께끼에서 다양한 인스턴스들의 결과들이 이러한 형태를 보였으므로 창의력에 기반하여 응답을 생성한 것이라고 간주될 수 있다. 따라서 전반적으로 LLM은 단순 태스크를 수행하는 태스크를 넘어서서 한층 더 추상적인 사고를 할 수 있으며, 추론 및 재치를 적절히 활용하는 정답들을 생성할 수 있다고 결론짓는다.

## 6. 결론

본 논문은 수수께끼 해결 능력을 바탕으로 LLM의 재치에 관한 고찰을 수행한다. 실험 결과 현재의 다양한 LLM 중 GPT-4가 가장 우수한 능력을 보였으며, 프롬프트 엔지니어링 과정에서 설명글을 추가하면서 동시에 데이터의 예시를 늘려서 제공하는 것이 성능 향상에 도움을 주는 것을 확인하였다. 또한 분석에서는 LLM이 주로 단어 기반의 말놀이 수수께끼가 아닌 특성을 활용한 수수께끼에서 상당한 성능을 보였다. 이는 영어권에서의 수수께끼가 특성 기반과 더욱 가깝다는 점을 들어 관련 데이터들을 영어 데이터 학습 시 더 많이 참고했을 것으로 분석했다. 이와 관련하여 LLM의 오류 유형 분석 시에도 특성 기반의 결과 중 일부가 비록 오답으로 간주되었으나 상당히 창의적이고 재치있는 답변을 생성했다는 점을 확인하였다. 이러한 결과들은 다양한 예시로부터 검증되었으며 반드시 학습으로 위 태스크에서 우수한 능력을 보인 것이 아닌 LLM이 창의력이나 추상적 사고, 재치와 추론 능력 등을 가지고 생성한 것으로 간주했다. 비록 LLM이 생각의 발산 측면에서 우수한 능력을 보이고 있으나 지나친 생각의 발산은 환각으로 이어지며 존재하지 않는 엔티티나 표현을 생성하는 결과가 있었다. 이에 환각을 발생시키지 않고 적절히 재치를 활용하여 창의력을 요구하는 태스크들을 수행할 필요가 있다고 결론짓는다.

## 감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2022-2018-0-01405). 또한 이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1A6A1A03045425). 또한 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-00368, 뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발).

## 참고문헌

[1] M. Lee, P. Liang, and Q. Yang, "Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities," *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–19, 2022.

[2] P. Sawicki, M. Grzes, F. Goes, D. Brown, M. Peeperkorn, K. Aisha, and P. Simona, "On the power of special-purpose gpt models to create and evaluate new poetry in old styles," 2023.

[3] G. Franceschelli and M. Musolesi, "On the creativity of large language models," *arXiv preprint arXiv:2304.00008*, 2023.

[4] J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, and E. Gibson, "A fine-grained comparison of pragmatic language understanding in humans and language models," *arXiv preprint arXiv:2212.06801*, 2022.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 27 730–27 744, 2022.

[6] OpenAI-Blog, "Chatgpt: Optimizing language models for dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[7] OpenAI, "Gpt-4 technical report," 2023.

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[10] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[11] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[12] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[13] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[14] R. Gubelmann, A.-l. Kalouli, C. Niklaus, and S. Handschuh, "When truth matters - addressing pragmatic categories in natural language inference (NLI) by large language models (LLMs)," *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pp. 24–39, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.starsem-1.4

[15] R. A. Georges and A. Dundes, "Toward a structural definition of the riddle," *The Journal of American Folklore*, Vol. 76, No. 300, pp. 111–118, 1963.

[16] Y. Zhang and X. Wan, "Birdqa: A bilingual dataset for question answering on tricky riddles," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 10, pp. 11 748–11 756, 2022.

[17] 신명선, "수수께끼의 메타언어적 성격에 대한 국어 교육적 고찰," *국어교육*, No. 110, pp. 4–90, 2003.

[18] 김열규, "수수께끼라는 언어 전략이 텍스트 상관성에 던지는 문제 몇 가지," *배달말*, Vol. 14, pp. 315–336, 1989.

[19] 안혜리, 황민아, and 최경순, "학령기 경계선지능아동의 수수께끼 유머 이해 능력," *특수교육논총*, Vol. 37, No. 4, pp. 27–41, 2021.

[20] B. Y. Lin, Z. Wu, Y. Yang, D.-H. Lee, and X. Ren, "Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge," *arXiv preprint arXiv:2101.00376*, 2021.