chatGPT의 Reasoning , Hallucination, Interactivity 중심으로 논문정리

<https://arxiv.org/pdf/2302.04023.pdf>

<읽기전에 기초개념 이해>

여기서 g p t 를 제로샷 학습으로 수행했다고 했는데,

-제로샷이란? -> 일반적으로 기계학습은 훈련데이터에 나타난 클래스나 레이블에 대한 예측을 수행하는 데 사용된다, 하지만 제로샷학습은 훈련데이터에 없는 클래스나 레이블에 대한 예측을 수행하는 것을 목표로 한다.

2.1 LLM을 확장하면 일반화 가능성이 향상되는 반면, 일반 LLM은 특정 응용 분야에서 부족할 수 있습니다

Thus, ChatGPT is able to preserve the knowledge

from pre-training and produce informative outputs

without access to external knowledge sources. ->이게 맞는말인가 외부지식에 접근없이 지식을 저장하는데 어떻게 hallucination이 일어나지? -> 웹문서로 사전트레이닝된 지식외의 질문이 나오면 혼동이 오는건가? 그렇다면 이걸안일어나게 하려면 어떻게 해야하지 , soft gpt역시 maual DB를 학습시켜서 그 DB기반의 informative outputs가 나오게 하는거 아닌가?

2.2

-

3.1.1

-NLP 자연어 처리에서 reasoning 추론이란 근본적인 개념이 뭘까?

일단 개념: 컴퓨터가 텍스트나 언어 데이터를 이용하여 논리적으로 추론하거나 추론 기반 작업을 수행하는 과정. 다양한 형태와 목적으로 사용된다.

QA, Sentiment Analysis, Information Retrieval, machine Translation 등등

4.1 연역추론은 "관측 또는 증거"에 기초하는 반면 귀납추론은 "관측의 진실"(즉, 반드시 참된 추론)에 기초합니다또 추론" – 연역은 전제에서 결론으로 이어지는 반면 유괴는 결론에서 가장 많은 것으로 나타납니다

4.1.1 추론의 종류 -> 연역적 추론 vs 귀납적 추론

연역적 추론은 보다 일반적인 것에 근거하여 구체적인 결론을 도출하는 과정을 포함합니다

전제. 반대로 귀납적 추론은 패턴의 구체적인 관찰, 처리를 포함합니다

결론적으로 연역적추론은, 일반적인 결론을 도출하기 위해 점점 더 추상적인 가상 유도 추론 더 적은 "guessing"을 요구 , 귀납적 추론은 하향식 상향식

뭔소리인지 몰라서 구글링 자료를 참조해봤다

Hallucination 원인,해결책 논문

<https://arxiv.org/pdf/2306.06085.pdf>

<일단 실험할땐 openai embeddings>

-abstract

We find that the use of context combined with embedded tags can successfully combat hallucinations

^embedded 태그와 결합된 컨텍스트의 사용 -> hallucination을 줄일수있다

We observed a significant reduction in overall hallucination when context was supplied along with question prompts for tested generative engines

^input값에 질문프롬프트+context의 사용

we evaluated how placing tags within contexts impacted model

responses and were able to eliminate hallucinations in responses with 98.88% effectiveness.

^ 컨택스트 내에 태그(이게 뭘까?) 를 배치하는 것이 무슨의미일까? [[이게 이 논문에서 설명하고자 하는 핵심 hallucination 방지법인것같다]]

-Introduction

In this paper, we aim to tackle the prevalent issue of hallucinations in LLMs, focusing on constraining

models to produce reliable and trustworthy output

^ 컨택스트 내 태그를 두는 것이 어떤방식으로든간 ‘제한’하는 방식이라고 한다

Selfcheckgpt ->이거 논문을 찾아봤는데 어떻게 이걸 계산하는지 나와있는것같다. 이해는잘안된다

^ hallucination을 방지/블랙박스 모델의 hallucination을 체크

-background

utilizing corpus-based

heuristics using relative word frequencies contribute to hallucination issues in generative models.

^관련단어빈출을 지표로 사용하는 휴리스틱 방식이 hallucination issue에 기여한다

-대규모의 텍스트 데이터 집합으로, 다양한 텍스트 소스에서 추출한 텍스트 문장, 문단, 문서의 모음 에다가 휴리스틱스 알고리즘 사용하여, 특정 말뭉치 내에서 단어가 나타내에서 단어가 나타나는 빈도를 다른단어나 단어그룹과 비교하여 분석하는 것. 예를 들어 특정 단어가 해당 주제와 더 높은 상대 빈도로 나타난다면, 해당 단어는 그 주제와 관련이 높다고 판단할수있고, 이것을 이용하여 이슈를 해결

Jailbreaking is a form of prompt engineering that seeks to modify or manipulate an AI model’s behavior or output beyond its intended capabilities.

^ Jailbreaking이라는 프롬프트 엔지니어링 형식이 LLM모델 개발자들이 걸어놓은 제한을 빠져나가는 방식으로 hallucination을 유발한다 <https://gist.github.com/coolaj86/6f4f7b30129b0251f61fa7baaa881516>

그렇다면 이 jailbreak형식으로 질문입력을 하면안되는것인가? context내용에 질문에대한 주관적 견해를 넣으면 jailbreak형식의 프롬프트이고, 이 프롬프트는 hallucination을 프롬프트를 작성한사람 의도대로 유발할수있다는 소리같다.

We will describe how the use of tagged context prompts can be used as an effective tool in reaching

this goal.

^ 태그된 context프롬프트가 핵심이다

-Method

included generating question sets, creating context prompts through summaries, verifying context prompts and questions, and performing experiments with different GPT models using context-based and context-free

^ 포함된 질문 세트 생성, 요약을 통한 컨텍스트 프롬프트 생성, 컨텍스트 프롬프트 확인, 질문 컨텍스트 기반 및 컨텍스트 프리를 사용하여 다양한 GPT 모델로 실험 수행

DataSetCreation

효과적인 질문 세트를 생성하기 위해서 context를 다양한 문서에서 효과적으로 요약해내야된다.

질문set 구성 -> input 질문 + context + 어떻게 답해라

문서요약에서 Gpt training data는 포함하지 않는다

?? 문서를 통해 질문이 생성됬다? , 근데 hallucination이 발생되는 과정은 질문을 입력하고 그 질문에 포함된 context나 여러 번의 질문답변 으로 학습된 llm모델이 여태까지의 답변이나 context외의 답을 하는건데, 왜 이런얘기를 하는거지?

Verification

1.상호 참조: 컨텍스트 프롬프트를 원본 소스?? -> 그러면 원본문서에서 컨택스트 프롬프트를 만들고 그와 관련된 질문을 만들어냈다는거냐? 와 상호 참조하여 해당 메시지를 확인

2. 관련성 평가: 생성된 질문이 context프롬프트와 관련이 있는지 확인

3. 일관성 검사: 질문에 대해 논리적 일관성과 언어적 정확성에 대해 검사. 이 단계를 통해 생성된 질문이 의미가 있고 문법적으로 타당한지 확인하였습니다,

생성 언어 모델의 능력에 대한 의미 있는 평가를 가능하게 합니다.

TagPlacement

위논문에서 context에 태그를 내장하는 프로그램을 개발했는데, 컨택스트…(source X) 컨택스트… 이런 형태로 설계된거 같고, 여기서 x는 고유식별자를 나타냄

+프롬프트에 “세부사항 제공 및 답변에 출처포함”을 끝부분에 추가해놓음

세부적인 절차

1.고유 태그 생성: 출처별 식별자 역할을 하기위해 고유 4자리의 숫자 생성

Ex) 소스 3626 -> 3626같은 태그는 소스?? 여기서 말하는 소스가 무슨소스지->참조한 위키디피 백과 문서이정도? 의 특정라인과 고유하게 일치 \*\*6쪽의 source material과 일치한다는데 이거에 대한 설명이 없다

2.태그를 context에 삽입: 프로그램에서 생성된 태그 숫자를 context에 통합, 언어모델이 이를 인식하는데의 문맥적인 힌트를 제공(그래 알겠는데, 이 숫자가 어떻게 힌트를 제공한다는거야? 내생각엔 만약 context가 참고한 문서의 유형이 사회/.과학쪽이면 3000~4000번 , 일반문서면 0~1000 이런식으로 생성하는것같음)

Experiment

다양한 문맥을 고려한 프롬프트 설계방식으로 context+tag 방식의 프롬프트 실험을 한다.

1: context를 포함하지않은 프롬프트 질문 input 인데 “자세한 정보를 제공하고 답변에는 출처를 포함하시오” 라는 문장을 질문에 추가하는 형식

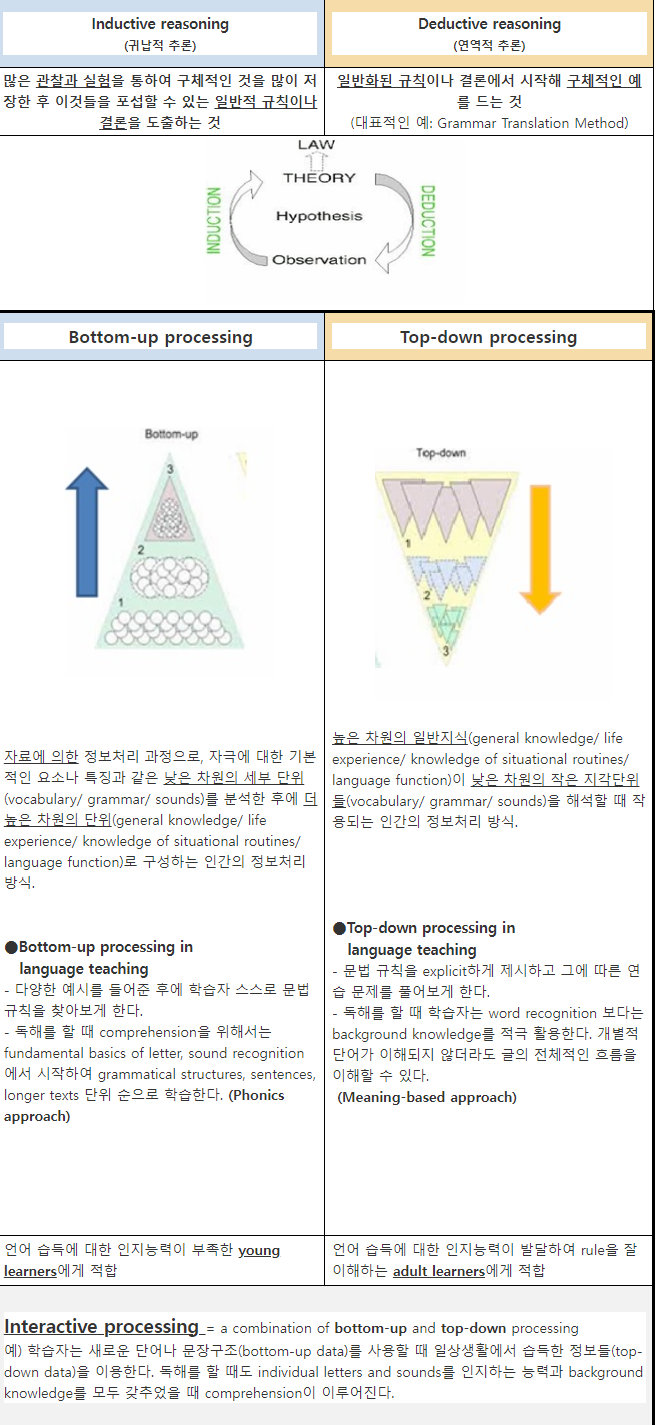
2: prompt에 사전생성된 tag를 포함한 context를 넣음

3: LLM에 프롬프트를 테스트해봄(대규모 LLM gpt-4, gpt-3.5 turbo)

Data Collection(<https://github.com/pgfeldman/KeywordExplorer/blob/main/keyword_explorer/experiments/contexts_1.json>)

위 깃허브에 나온 것 처럼 Experiment에 나온 과정의 결과를 Json형식으로 담아서 나타내고 있다.

<Context +자세한 정보를 주고 출처를 제공하시오 + 질문>



표를 통해 어떻게 추론이 이루어지는지 예시를 참고해보자.. -> 연역추론 -> 그냥 입력문의 나열 vs 프롬프트 엔지니어링(+Based on the given facts above, do a reasonable inference on this question using deductive reasoning:)

귀납추론 ->연역추론과는 다르게 프롬프트에 그냥 정보의 나열만 했을땐,

연역,귀납 방법을 떠나서 프롬프트에 ‘합리적인 추론을 하라는’ 문장을 입력한 것이 더 답변이 잘나왔다.

4.2