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>

-임베딩 사용-

<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(prompt) 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?~~

Prompt Verification

1.상호 참조: 컨텍스트 프롬프트를 여러가지 문서를 참조해서 만든다., 그리고 이 context를 참조할수있는 질문셋을 만든다.

2. 관련성 평가: 생성된 질문이 context프롬프트와 관련이 있는지 확인

3. 일관성 검사: 질문이 논리적 일관성과 언어적 정확성에 대해 검사

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 , 공학문서면 0~1000 이런식으로 생성하는것같음 // 우리가 적용해 볼수있는 부분: context를 만드는 DB에서 지식별로 숫자를 부여하는 방식 )

Experiment

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

case1: context를 포함하지않은 프롬프트 질문 input 인데 “자세한 정보를 제공하고 답변에는 출처를 포함하시오” 라는 문장을 질문에 추가하는 형식

case2: 1+prompt에 사전생성된 tag를 포함한 context를 넣음

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.(근데 여기에 context에 포함된 source라는게 없어?????)

<Context +자세한 정보를 주고 출처를 제공하시오 + 질문>

-그래서 3430개의 프롬프트를 LLM을 돌려서 응답을 받아냈고 이것을

응답유형별로 나눴는데

1715개의 응답은 컨텍스트와는 전혀 관계가 없고, 생성모델 고유의 지식에만 의존하고있다.(50%)

245개의 응답은 컨텍스트와 밀접한 관련이 있었다(7%)

1470개의 프롬프트-응답 쌍 에서 컨택스트 내의 질문이 문맥과 주제가 맞지않았다 라는 응답이 나왔다.

~~\*\*근데 prompt에 context랑 질문이 있는데, 이걸 LLM에 넣어서 응답을 뽑아내고, 이게 Hallucination인지 아닌지 판단하는 것은 응답이 context와는 동떨어진 내용이라던가, 질문의 형식이나 내용과는 거리가 먼 내용인지 아닌지 판단하는 것일텐데, 왜 자꾸 prompt-response pair에서 질문에대한 이야기를 하는거지????~~

Result

여태까지 이 논문에서 설명했던 task들 흐름에대한 실험결과에 대한 총정리를 하고있다

\*\*Prompt에 context가 없는 프롬프트-응답 쌍의 환각의 발생조사

* 위논문에서는 prompt가 context없이 구성될경우의 응답쌍에 대해 hallucination을 식별하는걸로 판단한것같다, 즉 프롬프트를 받은 LLM이 추가적인 context지침 없이 고유지식에 의존하는 response를 낼 때 hallucination이라고 정의했다 (근데 이건 이미 아는 내용이다..)

\*\*질문프롬프트에 context가 포함된 경우의 응답의 환각의 발생조사

* 확실하게 hallucination 생성이 크게 감소하였다고 했다.
* 원래 LLM 모델들은 질문과 context가 맥락상 일치하지 않아도 context를 참조하여서 응답을 도출하는경우가 많앗는데, 위 논문의 경우( 컨택스트에 고유 식별태그를 삽입하여 프롬프트를 구성 + 질문뒤에 “context를 참고해서 대답해주세요”라는 문장 삽입) 프롬프트에 질문과는 맥락상 관련이 없는 context가 포함되는경우 “ 제공된 컨텍스트가 질문과 일치하지 않습니다” 라는 응답이 나온다 [근데 이건 softgpt의경우에도 이렇게 나오지 않나?]
* 그래서 위 결과들로 아에 질문과 컨텍스트가 맥락이 다를경우 LLM이 질문에 허구의 지식이나 LLM자체지식을 참고하여 대답하는 Hallucination이 일어날 가능성이 급격히 감소하는걸 시사한다,

\*\* context 내에 태그를 배치하는 것이 LLM 에 넣었을 때 응답이 어떻게 나오는지

* Context 내에 태그를 배치하면 제대로된 응답에서는 context 와 일치하는 태그가 응답에 포함되어서 나오는데, 응답소스로서 일치 하는 태그“known good 소스 태그

4.3 Effects of Tag Placement in Context

Response가 제공된 source(위키디피아,나무위키 등의 지식DB)에 대해 하나 이상의 일치하는 태그가 있는 경우 “correct”로 간주 하였다

이런 기준으로 LLM에 태그가 포함된 244개의 컨택스트 프롬프트를 넣어 응답받는 방식으로 테스트를 해보았는데, 1개를 제외하고는 context의 태그를 참조하여 반환하였다.

This suggests that introducing relevant source tags within the context can significantly improve response accuracy by encouraging the generative language model to use and reference these known good sources.

이는 컨텍스트 내에 관련 소스 태그를 도입하면 생성 언어 모델이 이러한 알려진 좋은 소스를 사용하고 참조하도록 장려함으로써 응답 정확도를 크게 향상시킬 수 있음을 시사합니다. ??

일단 이 1개에 대해 설명하는데

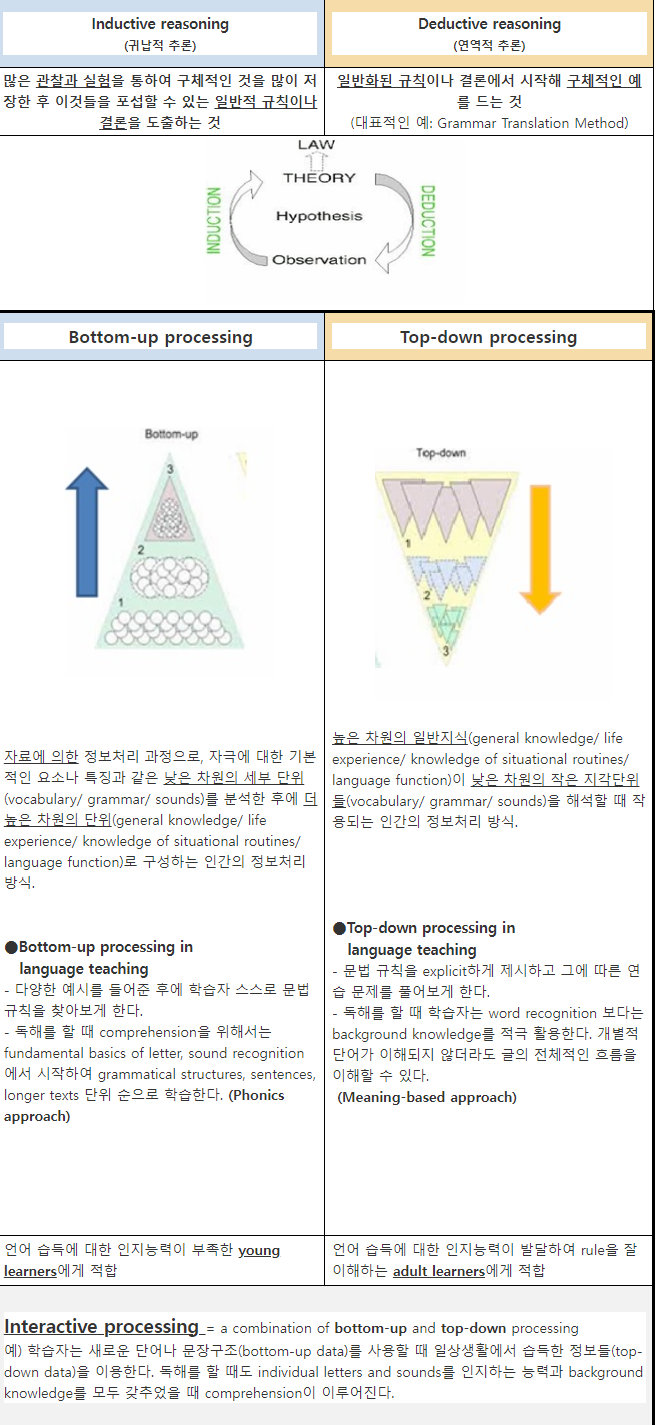
이유가 2가지가 있다

1.단일 컨텍스트 일경우

해소되지않은 질문: chatgpt같은 경우 입력창에 context가 아닌 질문만 입력하면, 답변이 형식과 질문의 의도에 맞게 잘나오는데, 이런경우 질문이 어떤 질문이 나올지 모르는데, context를 그에 따라서 어떻게 설정하는지, 아니면 질문 답변 쌍을 훈련을 시켜서 질문만 넣으면 답이 나오는건가? // 그러니까 context가 미리 결정되어있지 않고 질문만 넣으면 context를 설정해서 답변이 잘나오게 하는 방법,

이논문을 우리gpt에는 어떻게 적용할수있을까

질문 입력->컨텍스트 결정(리트리버 사용(이 질문에 맞는 컨텍스트를 매뉴얼 DB나 질답 데이터 셋에서 검색할때, 유사도를 질문과 제일 비슷한 후보들의 유사도를 이용하는대신, DB에 태그를 부여해서 질문과 비슷한 내용에 해당하는 DB에 태그)->프롬프트 완성->질답 데이터셋과 매뉴얼 DB가 pretrained 된 언어모델(Bert,gpt,vicuna)에 입력-> 추론->디코딩 및 답변출력 이런 형식으로 나오는게 맞는지



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