chatGPT의 Reasoning , Hallucination, Interactivity 중심으로 논문정리

<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

<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

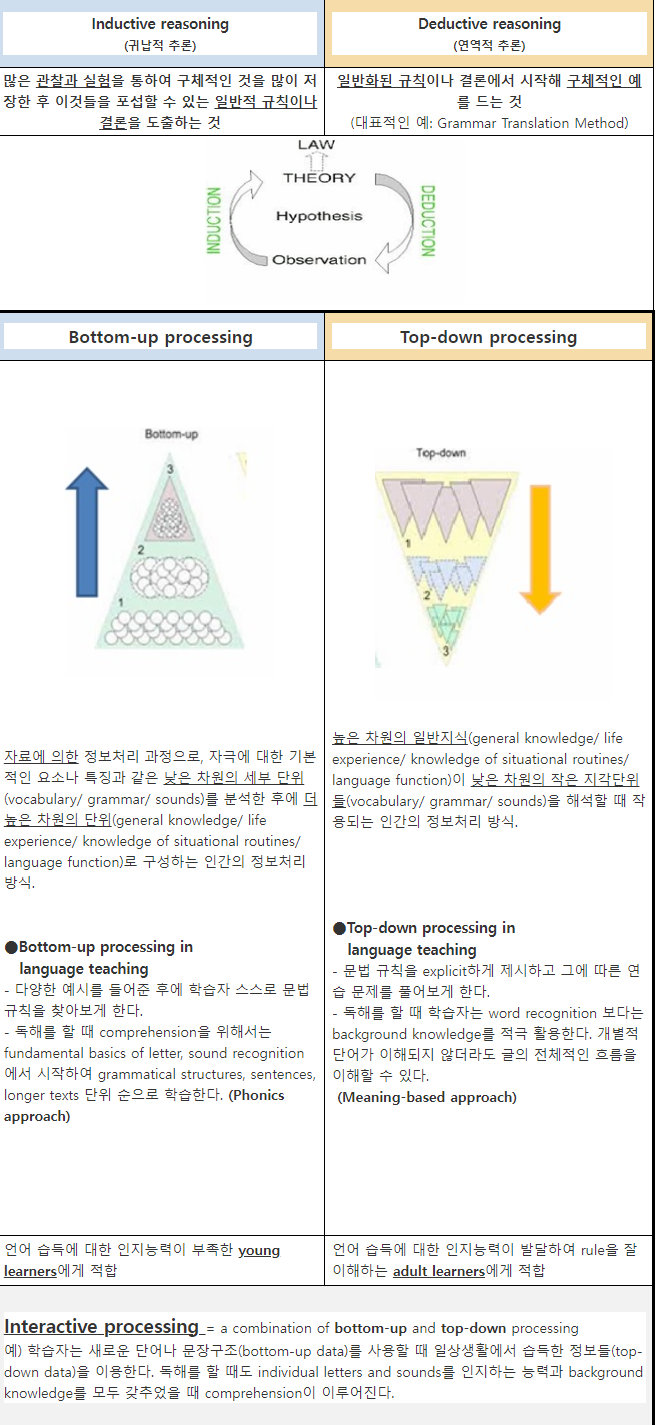
We observed a significant reduction in overall hallucination when context was supplied along with question prompts for tested generative engines

^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



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