HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models

Junyi Li[1,3,4]*, **Xiaoxue Cheng**[1]*, **Wayne Xin Zhao**[1,4]†, **Jian-Yun Nie**[3] and **Ji-Rong Wen**[1,2,4]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China
[3]DIRO, Université de Montréal
[4]Beijing Key Laboratory of Big Data Management and Analysis Methods
lijunyi@ruc.edu.cn chengxiaoxue3@gmail.com batmanfly@gmail.com

발표자: 임정우

Introduction

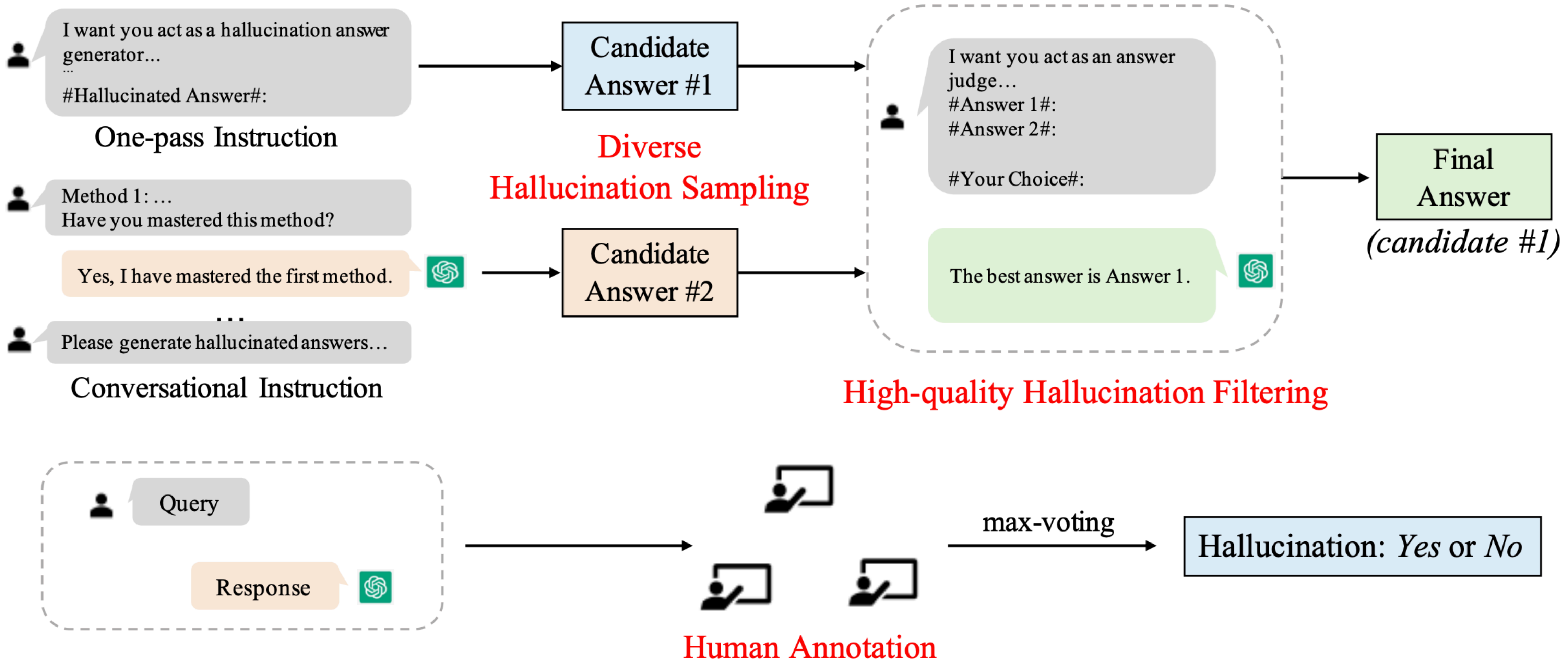
User Query	Retrieve the oldest photo of a cat
ChatGPT	There is no definitive answer to this question as "the oldest photo of a cat" is subjective and varies based on different sources and definitions. However, **one candidate for the oldest cat photo is the daguerreotype "Cat with Blue Ribbon," taken by photographer Antoine Claudet in 1840.** The photo depicts a cat sitting on a table, gazing directly at the camera, and wearing a blue ribbon tied around its neck. This photo is notable for being one of the earliest examples of animal photography ever taken, and it also showcases the intricate details and sharp resolution that can be achieved with the daguerreotype process.

Table 1: An example showing that the response generated by ChatGPT is likely to contain some information that looks plausible but is factually incorrect. The **green** span indicates the hallucinated content that cannot be verified by existing facts.

- Hallucination이 발생하는 현재의 LLMs
- LLM 이 얼마나, 어느 정도로, 어떤 종류의 Hallucination을 발생시키는지 확인할 수 있는 Benchmark가 중요함

https://arxiv.org/pdf/2305.11747.pdf

HaLuEval Benchmark



https://arxiv.org/pdf/2305.11747.pdf

HaLuEval Benchmark

- ChatGPT

User Query	Retrieve the oldest photo of a cat
ChatGPT	There is no definitive answer to this question as “the oldest photo of a cat” is subjective [...] However, one candidate for the oldest cat photo is the daguerreotype “Cat with Blue Ribbon” taken by photographer Antoine Claudet in 1840. The photo depicts a cat sitting on a table, gazing directly at the camera, and wearing a blue ribbon tied around its neck. [...]
Hallucination	Yes
Fragments	the oldest cat photo is the daguerreotype “Cat with Blue Ribbon” taken by photographer Antoine Claudet in 1840.

- Human Labeler

Question	In what political party was the man who officially opened Royal Spa Centre in 1972?
Right Answer	Conservative
Hallucinated Answer	Labour Party

https://arxiv.org/pdf/2305.11747.pdf

Experiments

- Experimental Setup
  - Models: GPT3-davinci, text-davinci-002, text-davinci-003, ChatGPT
  - Tasks: Summarization, QA, Dialogue, General

Experiments

- Hallucination Recognition

Models	QA	Dialogue	Summa.	General
GPT-3 (davinci)	49.21	50.02	51.23	77.54
text-davinci-002	60.05	60.81	47.77	87.60
text-davinci-003	49.65	68.37	48.07	87.54
ChatGPT	62.59	72.40	58.53	86.22

Table 6: Accuracy (%) of evaluation models to classify whether the sample contains hallucinated contents in HaluEval benchmark.

1. LLM 들은 텍스트에 내재된 Hallucination을 잘 인지하지 못함
2. LLM들은 특히 정답과 비슷한 형태를 띄지만 작은 범위의 키워드가 틀릴 경우를 잘 인지하지 못함
   (i.e., 정답: U.S. Highway 60 vs. 잘못된 답: U.S. Highway 70)

https://arxiv.org/pdf/2305.11747.pdf

Experiments

- Hallucination Recognition

Tasks	#Failed	Type-I	Type-II	Type-III	Type-IV
QA	3109	1559	245	278	1027
Dialogue	891	465	344	82	-
Summa.	3868	3106	705	57	-

Table 7: Number of failed classified samples by Chat-GPT for each hallucination pattern.

Type1: Comprehension
Type2: factualness
Type3: Specificity
Type4: Inference

https://arxiv.org/pdf/2305.11747.pdf

Experiments

Error Type	Question and GroundTruth	Responses
Comprehension	What is a rare breed of dog that was derived as a variant of Rat Terrier, Shiloh Shepherd dog or American Hairless Terrier? *(American Hairless Terrier)*	One rare breed of dog that was derived as a variant of Rat Terrier, Shiloh Shepherd dog or American Hairless Terrier is the Teddy Roosevelt Terrier.
Factualness	Are the New Orleans Outfall Canals the same length as the Augusta Canal? *(yes)*	No, the New Orleans Outfall Canals and the Augusta Canal are not the same length... the Orleans Canal is approximately 3.6 miles (5.8 kilometers) long ... The Augusta Canal ... is approximately 7 miles (11.3 kilometers) long
Specificity	1. What profession do Kçbç Abe and Agatha Christie share? *(playwright)* 2. What genre do Superheaven and Oceansize belong to? *(rock)*	1. Author 2. Superheaven and Oceansize are not of the same genre. Superheaven ... play a style of alternative rock that draws heavily from 90s grunge and emo. Oceansize ... was a British progressive rock band that was active from 1998 to 2011.
Inference	1. Which band has more members, Muse or The Raconteurs? *(The Raconteurs)* 2. Which is currently more valuable, Temagami-Lorrain Mine or Meadowbank Gold Mine? *(Meadowbank Gold Mine)*	1. Muse has more members than The Raconteurs. Muse is a British rock band with three members ... The Raconteurs ... Therefore, The Raconteurs has four members. 2. I cannot provide the current valuation of Temagami-Lorrain Mine or Meadowbank Gold Mine. Meadowbank Gold Mine... was producing gold at a rate of approximately 220,000 ounces per year. On the other hand, Temagami-Lorrain Mine is a historic iron mine located in Ontario, Canada that has been inactive for many years. It is not currently producing any minerals.

Table 1: Examples of Four Types of Errors (text in brackets represents the groundtruth, text in blue indicates the entities compared and their information, text in red indicates where the model (GPT-3.5) makes mistakes).

https://arxiv.org/pdf/2304.10513.pdf

Experiments

- Hallucination Recognition

Tasks	#Failed	Type-I	Type-II	Type-III	Type-IV
QA	3109	1559	245	278	1027
Dialogue	891	465	344	82	-
Summa.	3868	3106	705	57	-

Table 7: Number of failed classified samples by Chat-GPT for each hallucination pattern.

Type1: Comprehension
Type2: factualness
Type3: Specificity
Type4: Inference

1. Comprehension Error 타입이 많은데, query의 의도를 잘못 해석하여 답한 케이스가 많다는 것임. 모든 QA, Dialogue 태스크에서는 entity를 단순히 바꾸는 정도의 대답을 하는 경우가 많았다고 함
2. Text summarization경우에는 source에 반영하여 요약하지 않고 기존에 학습한 지식을 반영하여 답을 한 경우가 많았음
3. 이는 LLM들이 factual hallucination 인지 아닌지 판단하고자할때 관련 지식을 잘 못 가져오는 것이라고 해석될 수 있음

Experiments

- Hallucination Recognition

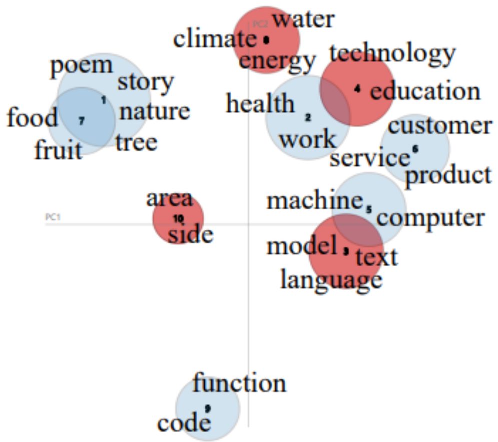


Figure 3: Topic distribution for general user queries and ChatGPT responses.

1. ChatGPT의 경우 Hallucination에 관한 질문을 할때 이를 잘 파악하는 topic이 있음
2. ChatGPT는 film, company, band등에 관한 토픽일 경우 잘 모르는 경우가 많음
3. 또한 아이러닉하게도 technology, language 등에서의 hallucination을 많이 구별하지 못했음

Experiments

- Improvement Strategies

Variants	QA	Dialogue	Summa.	General
ChatGPT	62.59	72.40	58.53	86.22
w/ Knowledge	76.83	73.80	-	-
w/ CoT	59.58	71.39	61.21	86.50
w/ Contrast	49.19	68.67	49.46	-

1. LLM에 관련된 지식을 넣어주면 hallucination이 훨씬 줄어듦
2. CoT는 QA나 Dialogue에서 LLM의 Hallucination을 구별하는 성능을 떨어트리는 모습을 보여주지만, Summarization에서는 성능이 올라감
3. Hallucination을 구별해주는 Sample을 미리 주면 더 헷갈려하는 상황이 발생하였음

Mitigating Language Model Hallucination with Interactive Question-Knowledge Alignment

Shuo Zhang[1,2] **Liangming Pan**[2] **Junzhou Zhao**[1] **William Yang Wang**[2]
[1]MoE KLINNS Lab, Xi'an Jiaotong University, P. R. China
[2]University of California, Santa Barbara, USA
{zs412082986@stu, junzhou.zhao@mail}.xjtu.edu.cn
{liangmingpan, wangwilliamyang}@ucsb.edu

발표자: 임정우

Introduction

- Hallucination이 발생하게 되면 retrieval-in-the-loop method 를 사용하여 해결하려는 접근이 있었음
  - commonly referred to as retrieval-augmented generation (RAG)
  - Knowledge Base를 참조하면서 evidence를 generation에 이용한다는 장점이 있음

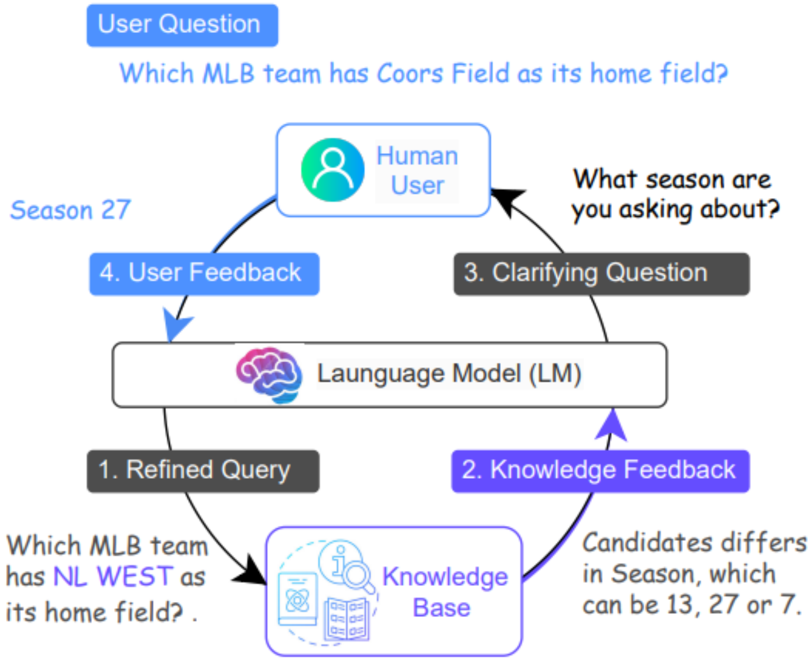
- 그러나, 아직도 hallucination 을 일관적으로 제거했다고 말할 수 없음.
  1. 특히, 생성된 텍스트가 찾아온 문서로부터 나온 것이 아닐 경우,
  2. 찾아온 문서를 아예 쓰지 않는 경우 등

- 그 이유는 user의 question과 stored knowledge와의 misalignment때문이라고 함
  - Knowledge base와 LM을 통합하는 과정에서 지속적인 간극이 있다!

https://arxiv.org/pdf/2305.13669.pdf

Method

- MixAlign
: a framework that interacts with both the user and the knowledge base to acquire clarifications on how the user's question relates to the stored evidence.



Retrieved Knowledge

Seasons	Team	Division
13	San Diego Padres	NL WEST
27	Colorado Rockies	NL WEST
7	San Francisco Giants	NL WEST

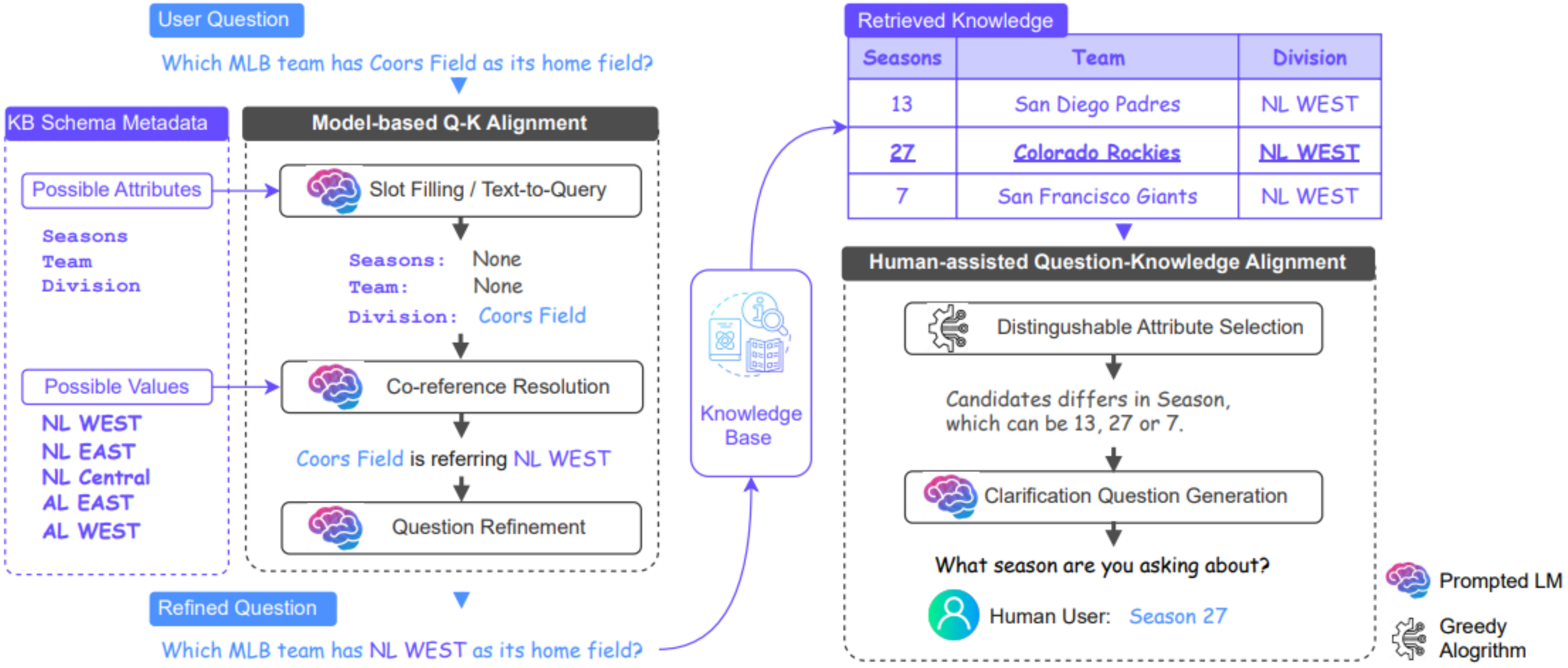
LM Answer

The Colorado Rockies has Coors Field as its home field in season 27.

https://arxiv.org/pdf/2305.13669.pdf

Method

- MixAlign
  : user의 question 표현과 저장된 knowledge의 간극을 줄이며, 불명확한 query일때, 이를 명확하게 해줌



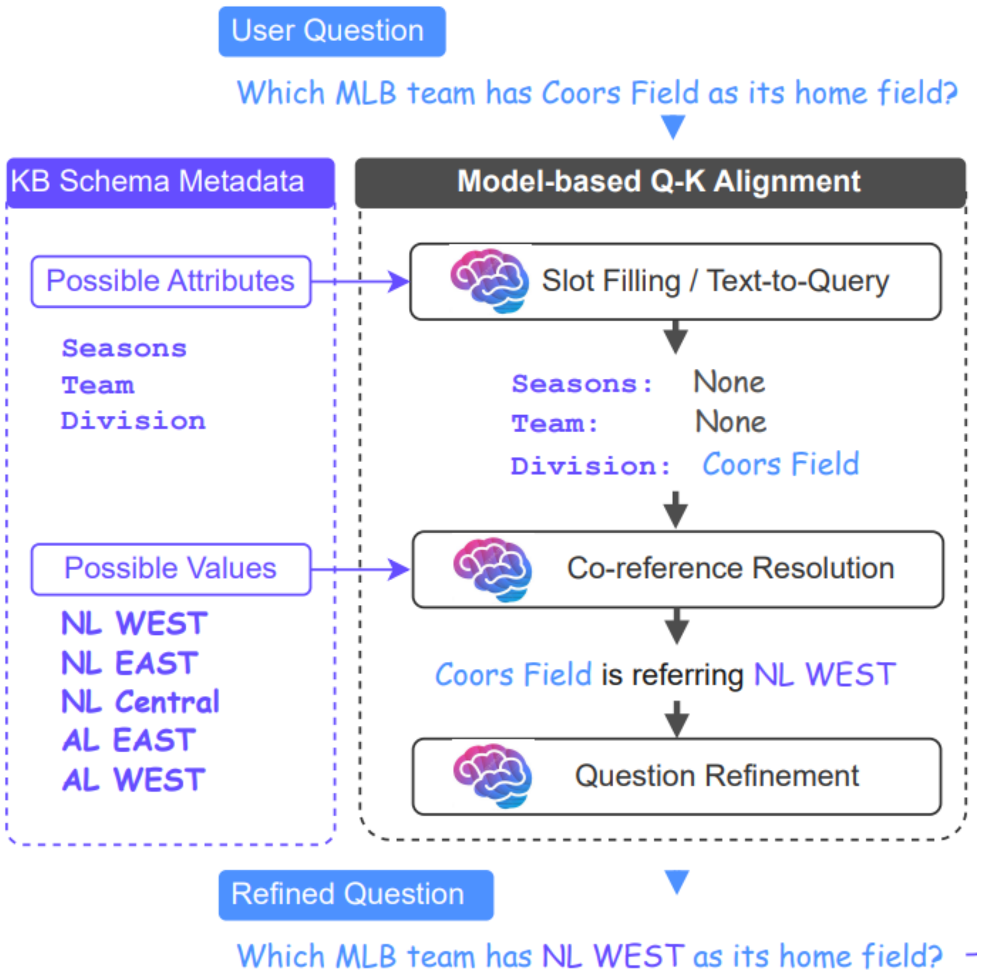
https://arxiv.org/pdf/2305.13669.pdf

Method

1. Model-based Question-Knowledge Alignment
    - Question을 먼저 입력받으면, 이 질문을 SQL query형태로 바꿈.

```
SELECT
    first_name
FROM
    employees
WHERE
    YEAR(hire_date) = 2000;
```

    - 그 다음, 데이터 베이스에 attribute name들에 채워진 value가 있는지 찾아봄
    - 찾아낸 value와 같거나 혹은 co-reference되는 value가 DB에 있는 경우를 찾음
    - 이를 기반으로 value를 수정하여 question-refining을 진행함



https://arxiv.org/pdf/2305.13669.pdf

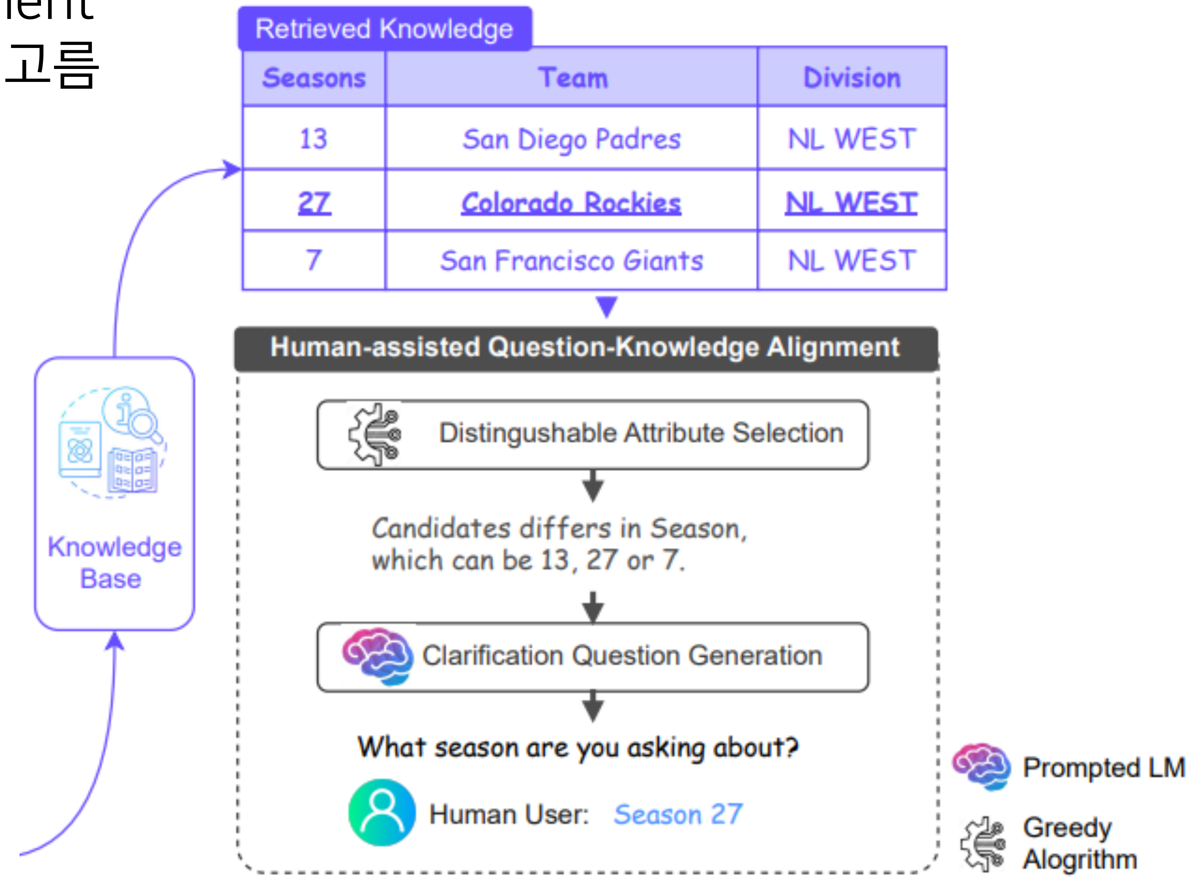
Method

2. Human-assisted Question-Knowledge Alignment
   - 가장 정답을 다르게 만들 수 있는 attribute을 고름

     > 이미 question에 나온 attribute, 그리고 ID attribute 은 제거

     > 그 다음, 가장 Unique value의 개수가 많은 attribute을 고름

   - 해당 attribute를 기준으로 question을 수정함



https://arxiv.org/pdf/2305.13669.pdf

Experiments

- Experimental Setup

  - Datasets: FuzzyQA 데이터셋 만듦 (HybridDialogue, MuSiQue 기반)
  - Models: GPT-3-text-davinci-003, RALM, CLAM
  - Metrics:
    1. Coverage: 생성한 답변에 gold answer가 있는지
    2. Hallucination: 정답과 질의에 없는 value가 생성한 답변에 있는지

Experiments

- Evaluation with Controlled Knowledge Groundings

  Q1: Do state-of-the-art Language Models (LMs) still hallucinate even when provided with accurate knowledge grounding?

  Q2: How does the presence of redundant irrelevant groundings impact LM hallucination?

  Q3: How does the alignment between a user's question and the stored knowledge affect LM hallucination? Is this alignment necessarily related to question complexity?

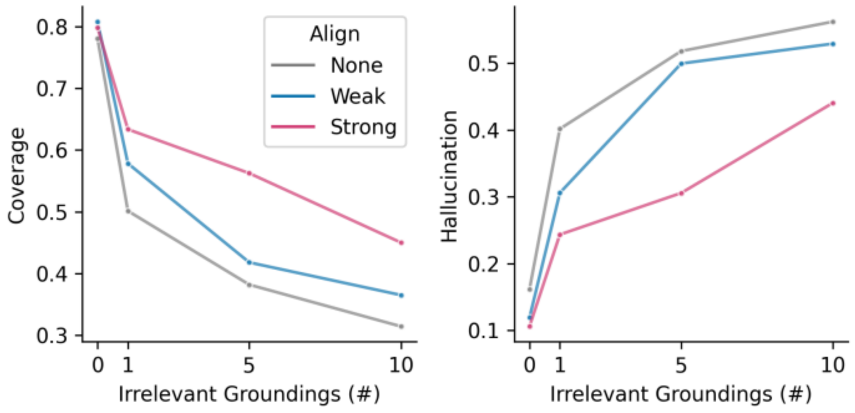


Figure 4: Automatic evaluation over coverage and hallucination for a varied number of irrelevant knowledge groundings, given different question-knowledge alignment degrees. The alignment is automatically measured using a slot-filling approach. In this method, we extract attributions from the gold knowledge grounding as slots and determine if these slots can be filled with information obtained from the user question.

Experiments

- Overall Evaluation

Method	Coverage ↑	Hallucination ↓
No Grounding		
Direct LM	9.42	82.68
CALM	17.73	83.21
Statistical Retrieval		
RALM	29.96	60.61
CALM	28.35	79.71
Model-based Question-Answer Alignment		
RALM	34.94	55.98
CALM	37.57	72.06
MixAlign (ours)	53.8	36.40

https://arxiv.org/pdf/2305.13669.pdf

PURR: Efficiently Editing Language Model Hallucinations by Denoising Language Model Corruptions

Anthony Chen[1*] **Panupong Pasupat**[2] **Sameer Singh**[1] **Hongrae Lee**[3] **Kelvin Guu**[1]
[1]University of California, Irvine [2]Google DeepMind [3]Google Search
{anthony.chen, sameer}@uci.edu {ppasupat, hrlee, kguu}@google.com

발표자: 임정우

Introduction

- Hallucination이 발생하게 되었을 때, 이를 Post-hoc Method 기반으로 수정하려는 접근들이 있었음
  - 이러한 접근은 어떠한 generation model을 썼어도 적용 가능하다는 장점이 있음

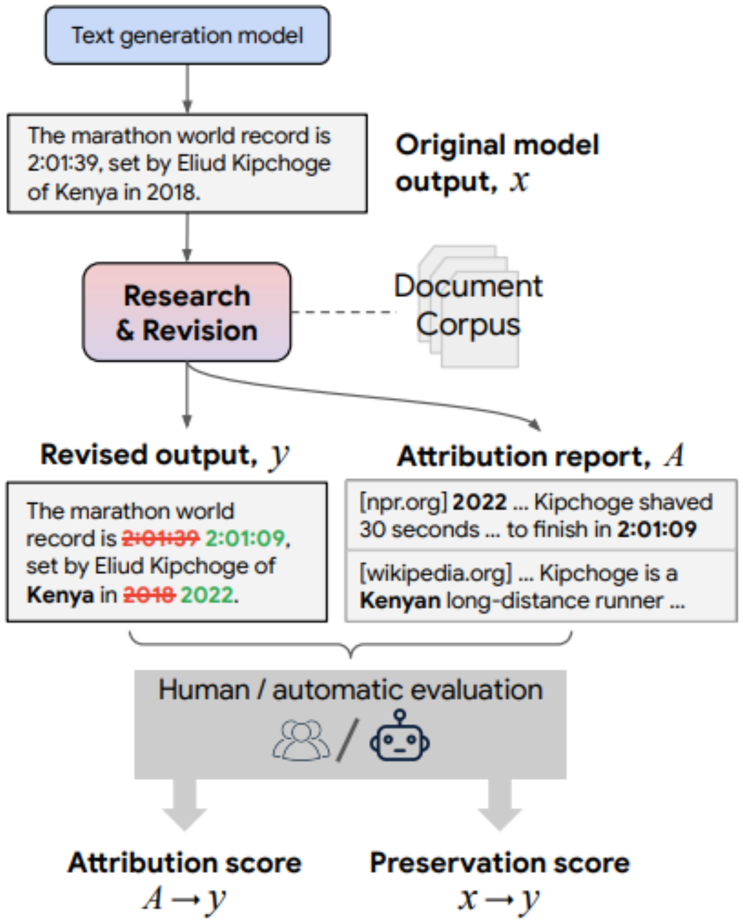
- LLM들은 이런식으로 editing 하기 위해서 단순히 few-shot prompting만 진행하면 된다는 장점과 그에 수반되는 막대한 비용이 있음

- 그에 반해, 작은 모델들은 fine-tuning 방법으로 비용을 최소화 할 수 있으나, 특정 도메인에 대한 데이터만 학습할 경우에 나타나는 한계가 있음 (generalization)

- 본 연구에서는, LLM과 작은 모델들을 이용하여 이 간극을 줄임

https://arxiv.org/pdf/2305.14908.pdf

Problem Formulation



1. Editing for Attribution

2. 먼저, Textual Statement x가 들어오면, 이를 기반으로 attribution report A를 생성함
(Attribution Report 는 evidence snippet들이고, x가 근거로 할 것 같은 진실된 정보들)

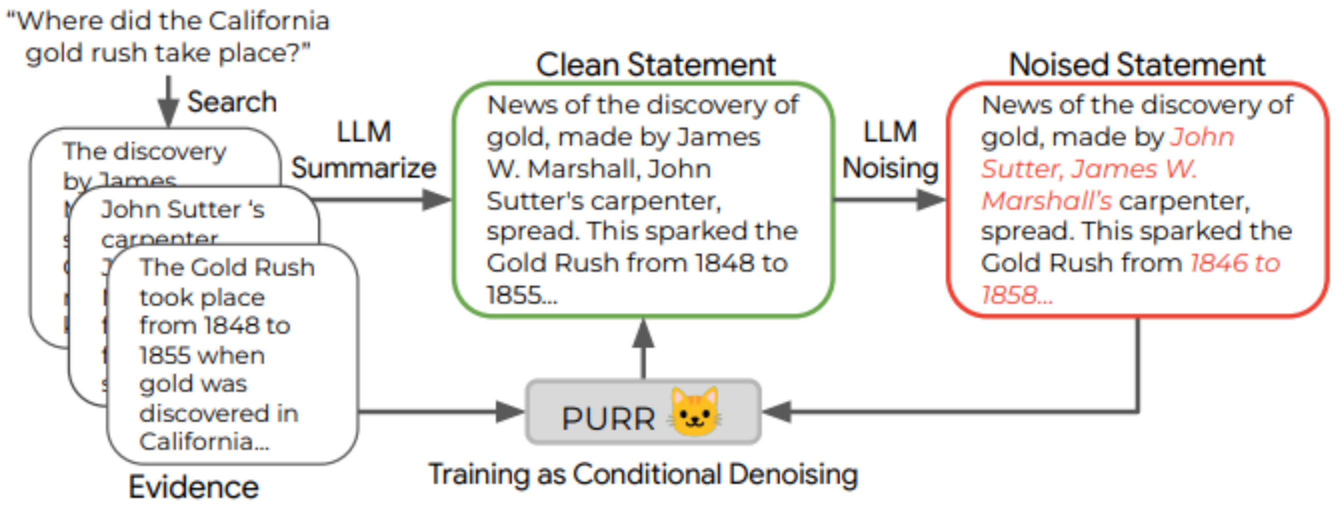
3. System은 그러면 A 를 기반하여 x 의 이상한 점을 수정한 y를 생성 해야하는 Task임

Attribution Score는 x와 y가 A에 얼마나 근거하고 있는지를 봄
Preservation Score는 x가 얼마나 y로 많이 바뀌었는지

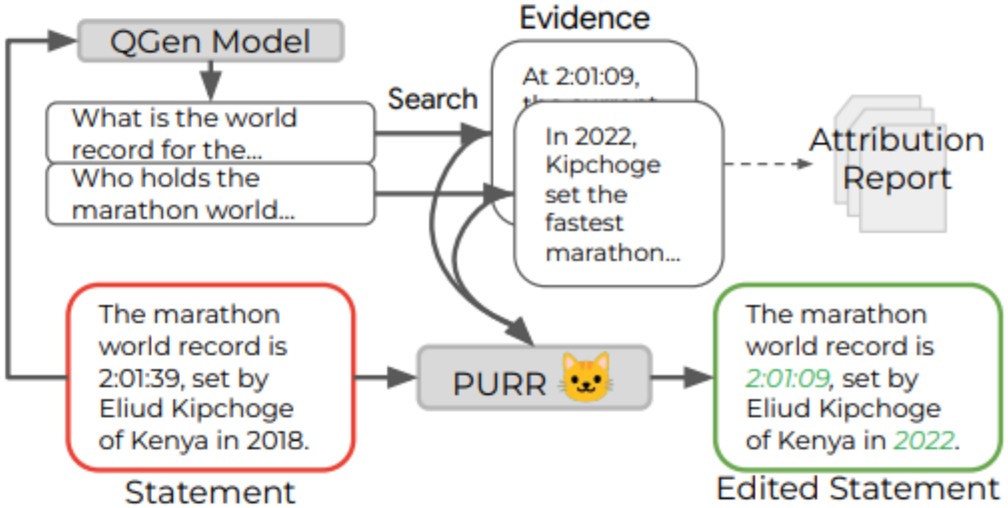
https://arxiv.org/pdf/2210.08726.pdf

Method

1. Petite Unsupervised Research and Revision (PURR)



(a) **Training PURR.** Given a seed query, we search for relevant evidence and summarize them into a claim which we corrupt. PURR is trained to denoise the corruption conditioned on the evidence.



(b) **Using PURR.** Given an ungrounded statement, we generate questions to search for relevant evidence which is then used to produce an edit.

Method

1. Petite Unsupervised Research and Revision (PURR)
   - Creating Training Data via Noising

     1) Question이 들어오면, 그 question과 관련된 웹페이지들의 passage들을 모두 모음
     2) 그 다음, cross-encoder 를 이용하여 가장 높은 점수를 가진 evidence들을 gold라고 가정함
     3) 나머지 passage들은 hard-negative라고 가정
     4) 그 다음, LLM이 gold evidence set을 요약하라고 prompting되고, 요약된 statement를 y라고 정의

   - Noising and Conditional Denoising
     1) LLM을 이용하여 y를 corrupt 시키라고 하면 이것이 statement x가 됨

https://arxiv.org/pdf/2305.14908.pdf

Method

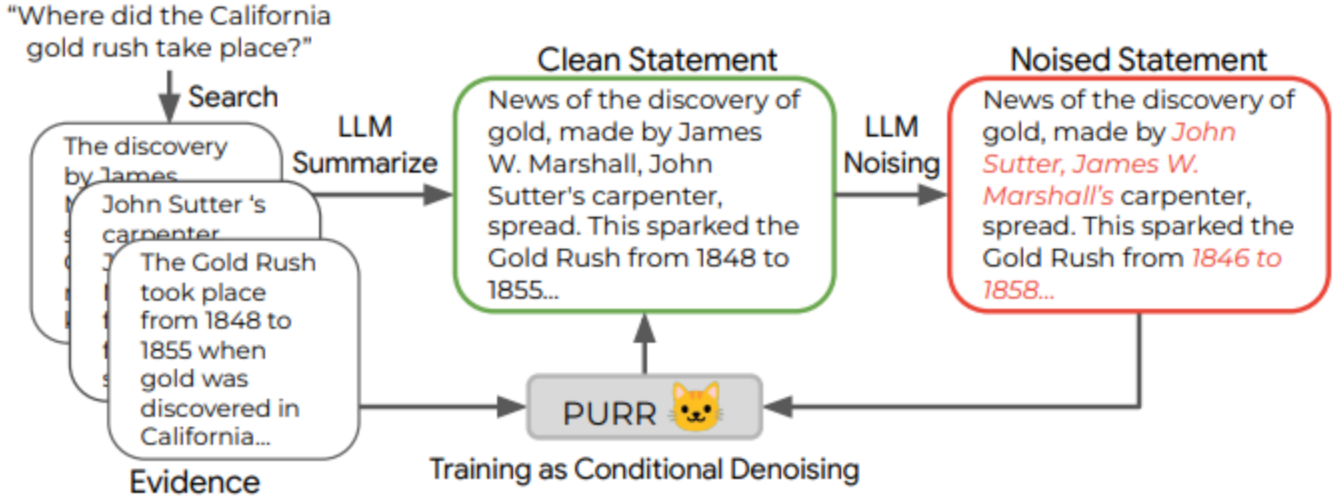
1. Petite Unsupervised Research and Revision (PURR)

q:	What is the neurological explanation for why people laugh when they're nervous or frightened?
E^+:	- A 2015 Yale study found people respond with a variety of emotions to strong outside stimuli… - Vilayanur Ramachandran states "We have nervous laughter because we want to make ourselves think what horrible thing we encountered isn't really as horrible as it appears"… - Stanley Milgram conducted one of the earliest studies about nervous laughter in the 1960s. His study revealed that people often laughed nervously in uncomfortable situations…
x/y:	Yale researchers in 2015 found people often respond to strong external stimuli with a variety of emotions, including ~~nervous laughter~~ ***anger***. ~~Stanley Milgram's~~ ***Vilayanur Ramachandran's*** 1960s study also observed this in uncomfortable situations. Neuroscientist ~~Vilayanur Ramachandran~~ ***Stanley Milgram*** theorizes that people laugh when….

Table 1: **Training Examples**. Our editing data covers a variety of domains and introduces challenging corruptions (*e.g.*, numerical, entity, and semantic role). q is the seed query, E^+ is the gold evidence set used to generate the clean statement, y is the clean statement and ***x is the corrupt statement***.

Method

1. Petite Unsupervised Research and Revision (PURR)



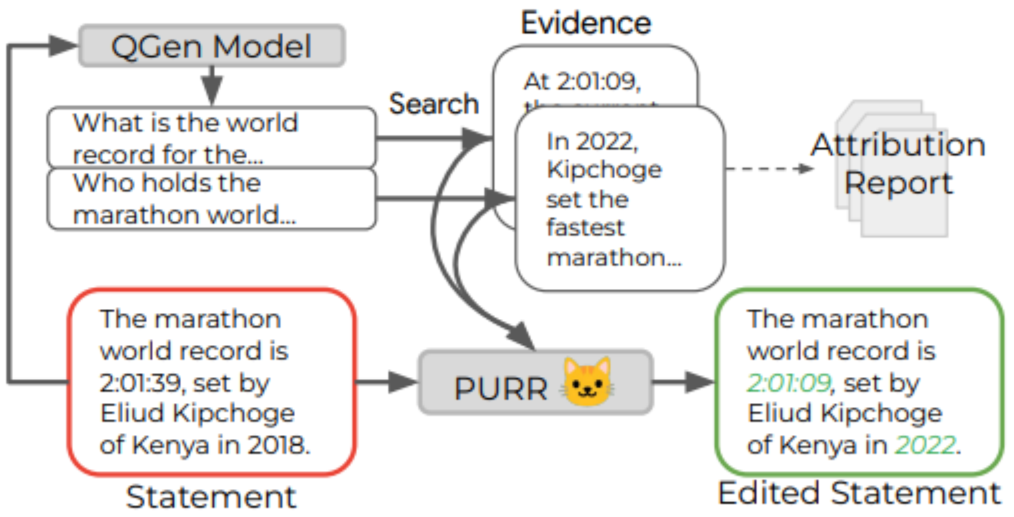
(a) **Training PURR.** Given a seed query, we search for relevant evidence and summarize them into a claim which we corrupt. PURR is trained to denoise the corruption conditioned on the evidence.

- PURR를 T5-large로 Finetuning

https://arxiv.org/pdf/2305.14908.pdf

Method

1. Petite Unsupervised Research and Revision (PURR)



(b) **Using PURR.** Given an ungrounded statement, we generate questions to search for relevant evidence which is then used to produce an edit.

- QGen Model은 Statement가 주어지면 Question 을 생성하는 Distillation 모델을 사용함

Experiments

- Primary Quantitative Results

Model	Attr. ($x \rightarrow y$)	Pres.	$F1_{AP}$
	PALM outputs on NQ		
EFEC	44.7 → **63.9**	39.6	48.5
RARR	44.7 → 53.8	89.6	67.2
PURR	44.8 → 59.8	**91.0**	**72.2**
	PALM outputs on SQA		
EFEC	37.2 → **58.2**	31.0	40.4
RARR	37.2 → 44.6	89.9	59.6
PURR	36.9 → 47.1	**92.0**	**62.3**
	LaMBDA outputs on QreCC		
EFEC	18.4 → **47.2**	39.0	42.7
RARR	18.4 → 28.7	80.1	42.2
PURR	16.8 → 33.0	**85.8**	**47.7**

Table 2: **Results on the *Editing for Attribution* task.** We report the attribution of the statement before and after editing, preservation after editing, and $F1_{AP}$ which combines attribution and preservation. Results are on LLM outputs on factoid question answering, long reasoning question answering, and dialog.

https://arxiv.org/pdf/2305.14908.pdf

Experiments

- Breaking Down the Numbers

  - **Huge Edit**: We say an edit is "huge" if preservation is low: $\text{Pres}_{(x,y)} < 0.5$.
  - **Bad Edit**: We say an edit is "bad" if the attribution after editing is lower than before: $\text{Attr}_{(y,A)} - \text{Attr}_{(x,A)} < -0.1$.
  - **Unnecessary Edit**: We say an edit is "unnecessary" if it is a bad edit and also $\text{Attr}_{(x,A)} > 0.9$. This means the editor made a poor edit when the attribution was already near perfect before editing.
  - **Good Edit**: We say an edit is "good" if attribution has significantly improved while preservation is high: $\text{Attr}_{(y,A)} - \text{Attr}_{(x,A)} > 0.3$ and $\text{Pres}_{(x,y)} > 0.7$.

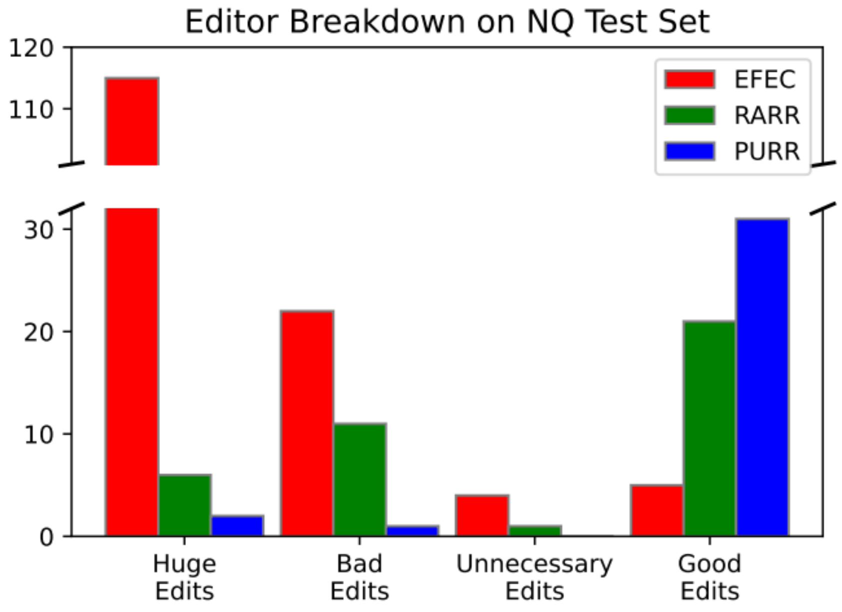


Figure 2: **Breakdown of edit types each editor makes** on the Natural Questions test set. EFEC makes huge edits while RARR sometimes over edits. PURR does a much better job at balancing attribution and preservation while rarely over-editing.

Experiments

- Qualitative Analysis

Correctly revising an entity/span.

x: The US stopped printing \$2 bills in 1966. The reason was that vending machines would not accept it. In 1976, the \$2 bill was reintroduced...

E: During that time, most goods and services were less than a dollar, making paper currency impractical. As need and use declined, the Federal Reserve stopped printing \$2 bills in 1966. In 1976, the Federal Reserve brought back the \$2-dollar bill.

y: The US stopped printing \$2 bills in 1966. The reason was that **most goods and services were less than a dollar.** In 1976, the \$2 bill was reintroduced...

$Pres_{(x,y)}$: 0.81 $\quad Attr_{(A,x)} \rightarrow Attr_{(A,y)}$: 0.61 $\rightarrow$ 0.84

Reasoning across multiple pieces of evidence

x: In 1899, the first driver's license was issued in France. The first license in the United States was issued in 1902 in Massachusetts.

E: - France introduced the first driving licences under the Paris Police Ordinance of 14 August 1893...
- Massachusetts and Missouri were the first states to require a license in 1903, but there was no test associated with it...

y: In **1893**, the first driver's license was issued in France. The first license in the United States was issued in **1903** in Massachusetts.

$Pres_{(x,y)}$: 0.90 $\quad Attr_{(A,x)} \rightarrow Attr_{(A,y)}$: 0.03 $\rightarrow$ 0.96

https://arxiv.org/pdf/2305.14908.pdf

Experiments

- Qualitative Analysis

Challenging distractors in evidence

x: The first orca put in captivity was in 1965. It was at Seattle, Washington's Seattle Aquarium. The orca's name was Wanda.

E: - On Nov. 18, 1961 a killer whale had been spotted and a collection crew from the **Marineland of the Pacific** scrambled to capture her. Never had an orca been captured alive, and Wanda was trying to avoid being the first...
- The first orca to be captured in 1961 was found off the coast of California and given the name Wanda.

y: The first orca put in captivity was in 1961. It was at Marineland of the Pacific. The orca's name was Wanda.

$Pres_{(x,y)}$: 0.77 $\quad$ $Attr_{(A,x)} \rightarrow Attr_{(A,y)}$: 0.33 $\rightarrow$ 0.77