DELUCIONQA: DETECTING HALLUCINATIONS IN DOMAIN-SPECIFIC QUESTION ANSWERING

MOBASHIR SADAT[1*], ZHENGYU ZHOU[2], LUKAS LANGE[2], JUN ARAKI[2], ARSALAN GUNDROO[2], BINGQING WANG[2], RAKESH R MENON[3*], MD RIZWAN PARVEZ[2], ZHE FENG[2]

[1]Computer Science, University of Illinois Chicago

[2]Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)

[3]UNC Chapel-Hill

[*]Work done during an internship at Bosch Research North America.

Motivation

- LLMs have a key weakness: *Hallucination*
  - Large Language Models (LLMs) are powerful,
    but they may hallucinate, i.e., generating non-factual content.
- Retrieval-Augmented LLMs **still hallucinate**
  - Hallucination can be relieved by leveraging information retrieval (IR) to provide additional context to LLMs, but it still happens from time to time (*See an example in Figure 1*)
- The problem is critical for question-answer (QA) applications requiring high reliability
  - A hallucinated answer delivered to user may raise significant liability concerns (e.g., vehicle damage, driver safety).

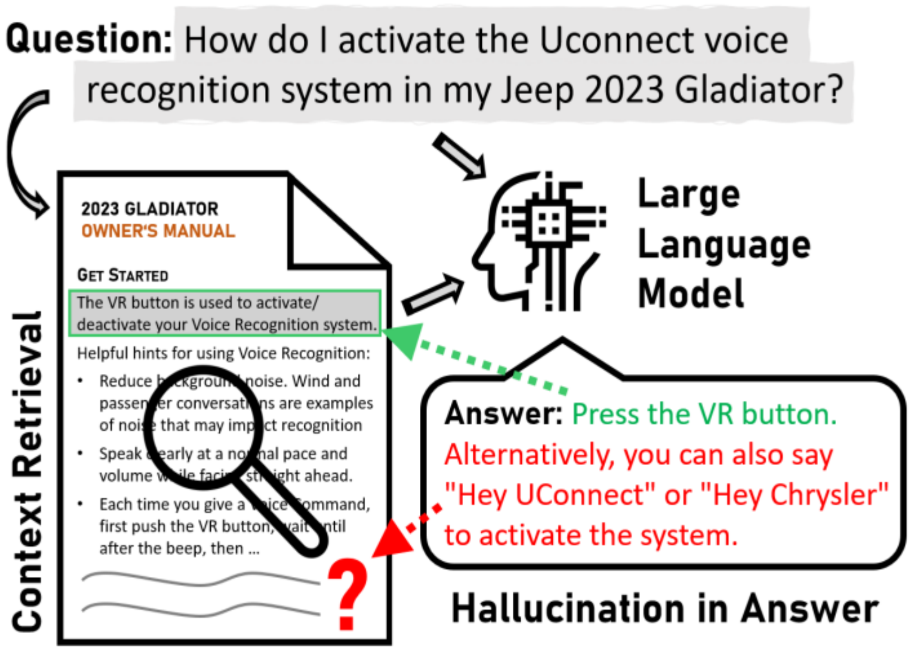

Figure 1: Hallucination in text generated by LLMs.

Contributions

- We facilitate the investigation of the hallucination phenomenon in retrieval-augmented LLMs for domain-specific applications with high reliability requirements, by:
  - Collecting and presenting a large QA dataset, named DelucionQA, for this study
    - Without loss of generality, the dataset adopts car-manual QA as a domain-specific representative task, and a state-of-the-art GPT model as a representative LLM.
    - Multiple retrieval methods are implemented for QA with retrieval-augmented LLM.
    - The occurrences of hallucination in answers are labeled by crowdsourcing workers
  - Proposing a set of hallucination detection approaches to serve as baselines for our dataset
    - The best performing baseline method shows a Macro F1 of only 71.09%, illustrating the challenging nature of the task.
  - Providing insights on the causes and types of hallucinations

Dataset Construction

1

Step 1: Question Generation

2

Step 2: Context Retrieval

3

Step 3: Answer Generation

4

Step 4: Manual Annotation

Dataset Construction

Step 1: Question Creation

- Automatic question generation by T5 model
  - A set of candidate questions are automatically generated based on the publicly available car-manual of Jeep 2023 Gladiator using a multi-task T5 model (Raffel et al., 2020)

- Manual refinement of the question set
  - The candidate questions are manually polished or filtered out if they are not realistic.
  - Additional relevant questions that are not generated by the T5 model are manually added.

Dataset Construction

Step 2: Context Retrieval

- IR techniques are used to retrieve relevant context for each question and <Question, Context> pairs are constructed.

- We use the following four IR methods to retrieve a variety of context for each question:
  - Sparse retrieval
  - Top-K ensemble retrieval
    - Top-1
    - Top-3
  - Adaptive ensemble retrieval

Dataset Construction

Step 3: Answer Generation

- We choose the OpenAI ChatGPT model as a representative LLM because it is arguably the most advanced, best performing, and widely accessible LLM at the time of writing our paper.
- ChatGPT is prompted with each <Question, Context> pair constructed in Step 2 to generate an answer to the question based on the provided context.
- This step results in a large number of <Question, Context, Answer> triples.

Dataset Construction

Step 4: Manual Annotation

- Amazon Mechanical Turk platform is used to assign labels for each sentence in the answer indicating whether it is (1) *supported, (2) conflicted,* or *(3) neither supported not conflicted* by the context in each triple.

- Based on the sentence level labels, a sample level label is assigned to each triple as follows:
  - Hallucinated: if any of the sentence level labels in the triple is *conflicted* or *neither supported not conflicted*.
  - Not Hallucinated: if all sentence level labels in the triple are *supported*.

Dataset Splits and Statistics

Split	#Ques	#Triples	#Hal	#Not Hal
TRAIN	513	1,151	392	759
DEV	100	216	94	122
TEST	300	671	252	419
TOTAL	913	2,038	738	1,300

Table 1: Number of unique questions, number of triples and label distribution in each split of DELUCIONQA. Here, Ques: Question and Hal: Hallucinated.

Hallucination Detection Approaches (Baselines)

Sim–cosine: based on sentence-level embedding similarity between the context and the answer.

Sim–overlap: based on sentence-level overlap between the context and the answer.

Sim-hybrid: a hybrid of Sim–cosine and Sim–overlap.

Keyword-match: based on the ratio of keywords match between the context and the answer.

Performance of Proposed Approaches

Method	Train			Dev			Test		
	Hal	N-Hal	Overall	Hal	N-Hal	Overall	Hal	N-Hal	Overall
SIM-COSINE	63.18	74.73	70.03	72.45	77.12	74.78	63.84	73.55	69.45
SIM-OVERLAP	68.47	82.72	75.59	73.51	80.16	76.84	63.89	78.28	71.09
SIM-HYBRID	68.73	83.16	75.94	73.51	80.16	76.84	63.33	78.29	70.81
KEYWORD-MATCH	30.25	77.47	53.86	31.58	69.57	50.57	31.23	74.31	52.77

Table 2: Class-wise F_1 scores (%) and overall Macro F_1 scores (%) of the baseline hallucination detection methods on the three splits of DELUCIONQA. Here, Hal: Hallucinated, N-Hal: Not Hallucinated.

Case Study

Based on our qualitative analysis, we identify the main reasons for different cases of hallucination:

- P1: using learned knowledge instead of context
- P2: unrelated information in retrieval results
- P1 and P2 can occurring jointly.

  Note: Whether or not an LLM hallucinates partially depends on how strongly it is biased towards the pre-learned knowledge.

Question	Answer
How do I activate the UConnect voice recognition system?	. . . Alternatively, you can also say **“Hey UConnect”** or **“Hey Chrysler”** to activate the system . . . . . . or say **“Hey UConnect”** followed by a voice command . . .
What does emission inspection check?	. . . The inspection **also checks for any issues that may cause the MIL to be illuminated during normal vehicle operation** . . . . . . The inspection also checks for **fluid level**, exhaust system, and **catalytic converter maintenance**.
How far can I use key fob to lock/unlock the door?	. . . in general, the range can be up to **several hundred feet** . . . . . . from distances up to 66 ft (20 m) using the key fob . . .

Table 3: Examples of answers generated for given questions, with hallucinated parts in **bold**.

Conclusion

- This work addresses hallucination in answers generated by LLMs in retrieval-augmented QA, focusing on a domain-specific scenario where high reliability is required.
- We present a dataset (DelucionQA) containing 2,038 samples, together with a number of baseline hallucination detection approaches, to facilitate research in this direction.
- Qualitative analysis is also conducted to provide insights into why hallucination occurs.
- While DelucionQA is constructed from a single representative domain (car manual) with one representative LLM (ChatGPT), we believe that the insights obtained and the approaches developed can be extended to other domains/LLMs as well.
- Our future work will include diversifying our study to multiple domains/LLMs, and developing more advanced hallucination detection/handling approaches.

Thank you!