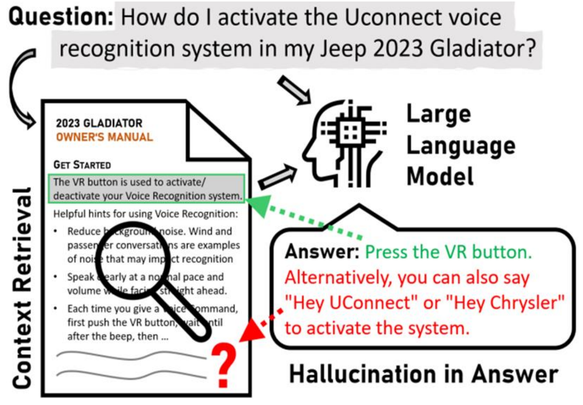
DelucionQA: Detecting Hallucinations in Domain-specific Question Answering

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R Menon, Md Rizwan Parvez, Zhe Feng

Motivation

- Large Language Models (LLMs) are powerful, but they have a key weakness: **Hallucination** (i.e., generating non-factual content).
- Retrieval-Augmented LLMs **still hallucinate.** The problem is critical for question-answer (QA) applications requiring high reliability



Contributions

- Construct/release a dataset, "**DelucionQA**", to facilitate hallucination research for retrieval-augmented LLM-based domain-specific QA.
  - Without loss of generality, car-manual QA (with high reliability needs) and ChatGPT are chosen as the representative domain and LLM, respectively
- Propose baseline hallucination detection methods
- Provide insights on causes/types of hallucinations

DelucionQA Construction & Statistics

1. Question Generation
2. Context Retrieval
3. Answer Generation
4. Manual Annotation

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.

Results

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

Conclusion

- We release a new dataset, together with baseline approaches and analyses, to facilitate the study of hallucination in retrieval-augmented QA applications with high reliability requirements.
- While DelucionQA is constructed for the car-manual domain with ChatGPT, the insights obtained and the approaches developed can be extended to other domains/LLMs as well.
- Future work will involve incorporating other domains/LLMs, and developing more advanced hallucination detection/handling approaches.

Contact: msadat3@uic.edu; Zhengyu.Zhou2@us.bosch.com

Dataset Link: https://github.com/boschresearch/DelucionQA