Team Elevate Proposal: Hallucination Hacker

Shreya Shah, Anusha Shaikh,
Noor Syed & Nabeeha Ahmed

Github link

Problem Statement

How can KPMG leverage the Hallucination Hacker Tool to detect and prevent adversarial hallucination exploitation in AI systems by categorizing outputs based on source credibility, supporting evidence and cross-verification, then flagging uncertain results and informing the user that accurate information was not found, thus preventing the hallucination from impacting the user?

Solution

Design an Python-based application that assesses the credibility of sources based on the a data set provided, thus preventing AI from generating hallucinations.

Process

Analyze

Prevent

Prototype Demo

Github link

Framework

```
# print(web_results)
cred = []
for result in web_results:
    rows = (my_conn.execute( query: 'SELECT rank FROM domains WHERE url = ?',

    if rows:
        cred.append(rows[0][0])
    else:
        cred.append(0)

# find the average value and if its greater than or equal to 3 say yay for
average = sum(cred) / len(cred)

print("The fact is: " + fact)
print("The credibility rating is: " + str(average))
if average >= 3:
    print("The information found appears to be credible.")
else:
    print(
```

Statement Input

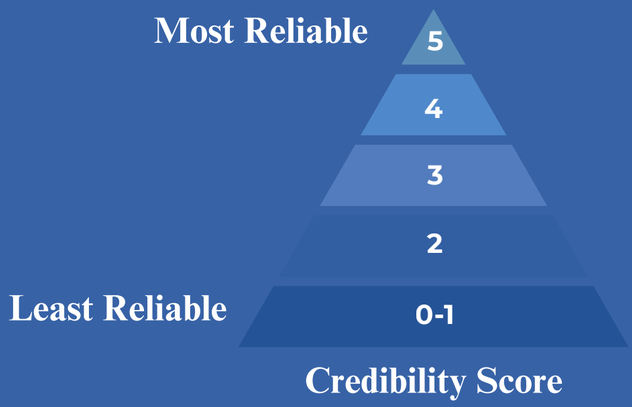
```
if __name__ == "__main__":
    fact1 = "eating vegetables is really important"
    fact2 = "Electroconvulsive therapy (ETC) is used in surgery"
    fact3 = "blueberries are peaches"
```

Statement Output

```
The fact is: Electroconvulsive therapy (ETC) is used in surgery
The credibility rating is: 0.6
The information appears to be found from a mix of less credible sites, a
The fact is: blueberries are peaches
The credibility rating is: 0.6
```

This demo showcases the ability for our tool to search the web and categorize the individual facts searched based on credibility through the Duck Duck Go search engine.

For example, we have entered the statement “Electroconvulsive therapy (ETC) is used in surgery”, this is an incorrect fact that resulted in a credibility of 0.6 rating, on a scale of 0-5, with 1 being least credible. The AI responde to the user with the statement “The information appears to be found from a mix of less credible sites, additional research would be beneficial”, therefore preventing external hallucinations.



Real-World Cases

Companies

- This tool enable companies to implement AI safely.
- Companies may input internal information based on their requirements
- Employees can validate information used in projects
- This tool can be altered to prevent phishing emails

Schools

- Implemented by schools and used by students as a resource to generate proper credible sources
- AI use can be monitored and can be trusted to provide factual information

Health

- Hallucinations produced by AI can lead to incorrect self diagnoses, this tool ensures valid responses
- Many people use AI for healthcare related questions, hallucination Hacker prevents the spread of mis-informed health trends

Objectives

01

Innovation & Originality:
Hallucination Hacker is a disruptive innovation, as it is the implementation of credibility checking to prevent AI generated hallucinations.

02

Technical Feasibility

Our tool is a realistic solution to keep up with the evolving developments in AI, to keep AI outputs under control.

03

Impact & Relevance

The tool targets the prevention of hallucinations produced by AI by verifying factual information based on credible sources. If the sources retrieved by the AI rank low in multiple categories, no false information is given.

04

Scalability & Applicability

The scalibility of this tool can easily be reached to other industries, platforms and enviornments. By rephrasing the database, it can be used by anyone who wants to retrieve credibility. Adaptable across cloud, on-prem, edge and hybrid.

05

User-Centric Design

The AI tool is intuitive and accessible as it can be altered based on any issue, there is no psycology as AI is trained to jsut the dataset while recieving a credibility score to the facts entered. Anyone can simply use this tool to fact-check and prevent hallucinations.