Meeting Note

Author: Henry Lin
Date: 2024/07/29 – 2024/08/02

Weekly Summary
LLM Hallucination Definition [1], Evaluation [2].

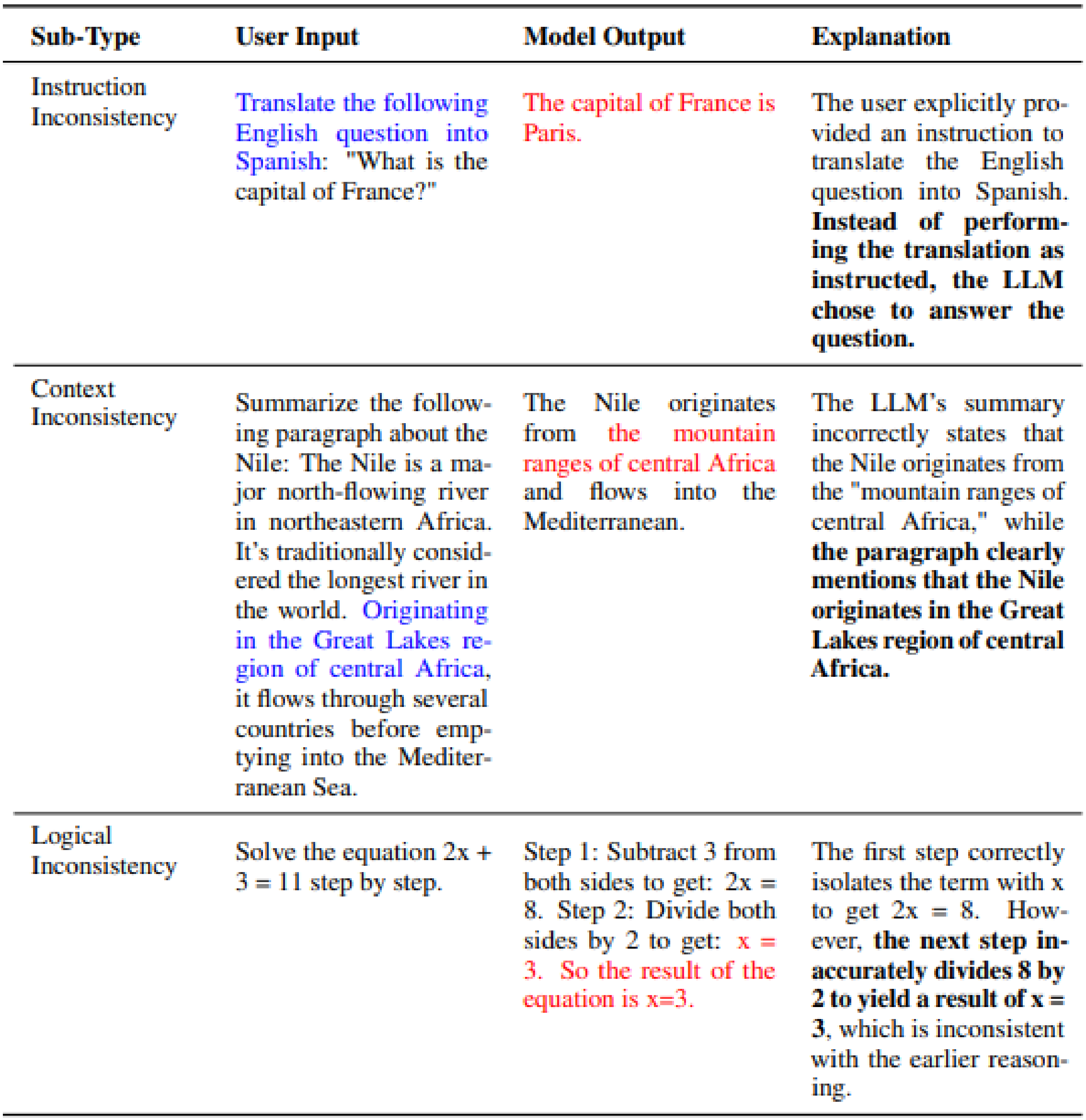
Plan for Next Week
LLM Mitigation.

Details
In this week, we focused on the definition and evaluation of LLM hallucination review; and in the next week, we will move on to the mitigation of hallucination.

In the study provided by Huang et al. [1], they describe the LLM’s hallucination in section 2.3 as follow: *the perception of any entity or even the absent in reality*, and divide the hallucination into two types: (1) Factuality Hallucination and (2) Faithfulness Hallucination. The factuality hallucination means that the generated content is factually inconsistent, for example, the LLM answers that the first human landed on the Moon is "Yuri Gagarin" (the correct answer should be "Neil Armstrong"), or it’s a made-up content such as "the origin of unicorns"; the latter is divided into three types further shown in Table 1.

Sun et al. [2] introduced an evaluation method to benchmark the hallucination in LLM by using the unanswerable mathematical questions. These unanswerable questions are divided into five classes: (1) Key Information Missing (2) Ambiguous Key Information (3) Unrealistic Conditions (4) Unrelated Object and (5) Question Missing. And each the answerable question in the benchmark has a definite answer. To evaluate and judge whether the output of LLMs is "answerable" or "unanswerable", a function is defined.
The f_{sim} in Figure 1 computes the similarity between a given sentence and a set of unanswerable template sentences. If the output contains mathematical expression or is similar to an unanswerable question, it will be labeled as "unanswerable", otherwise, it will be labeled as "answerable".

Apart from the previous method, Huang et al. [1] describe several more general methods to detect the hallucination. The following are the methods for Factuality Hallucination Detection and Faithfulness Hallucination Detection respectively.



Sub-Type	User Input	Model Output	Explanation
Instruction Inconsistency	Translate the following English question into Spanish: "What is the capital of France?"	The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. **Instead of performing the translation as instructed, the LLM chose to answer the question.**
Context Inconsistency	Summarize the following paragraph about the Nile: The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea.	The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The LLM's summary incorrectly states that the Nile originates from the "mountain ranges of central Africa," while **the paragraph clearly mentions that the Nile originates in the Great Lakes region of central Africa.**
Logical Inconsistency	Solve the equation 2x + 3 = 11 step by step.	Step 1: Subtract 3 from both sides to get: 2x = 8. Step 2: Divide both sides by 2 to get: x = 3. So the result of the equation is x=3.	The first step correctly isolates the term with x to get 2x = 8. However, **the next step inaccurately divides 8 by 2 to yield a result of x = 3**, which is inconsistent with the earlier reasoning.

Table 1. Three classes of faithfulness hallucination.

The first method is "Retrieval External Facts", once the question is given to the LLM and obtain an answer from it, the discriminator will use the same query to retrieval the real-world source, such as web source (e.g. Wikipedia), and check if the output from LLM is consistent with the real-world data. For faithfulness assessment, there are five types of metrics shown in the Figure 2.

Algorithm 1 Answerability Evaluation

```
Input: Generated text $v$ of a question by LLM
Output: Answerable or not
$S \leftarrow f_{\text{sim}}(v, u_i)$
if $\max(S) \geq T$ then
    return False
end if
$T \leftarrow \text{TokenizeText}(v)$
$T' \leftarrow \text{RemoveCommonVocabulary}(T)$
$v' \leftarrow \text{RemoveWhitespace}(T')$
if $\text{ContainsExpression}(v')$ then
    return False
end if
return True
```

Figure 1. The evaluation algorithm for LLM hallucination.

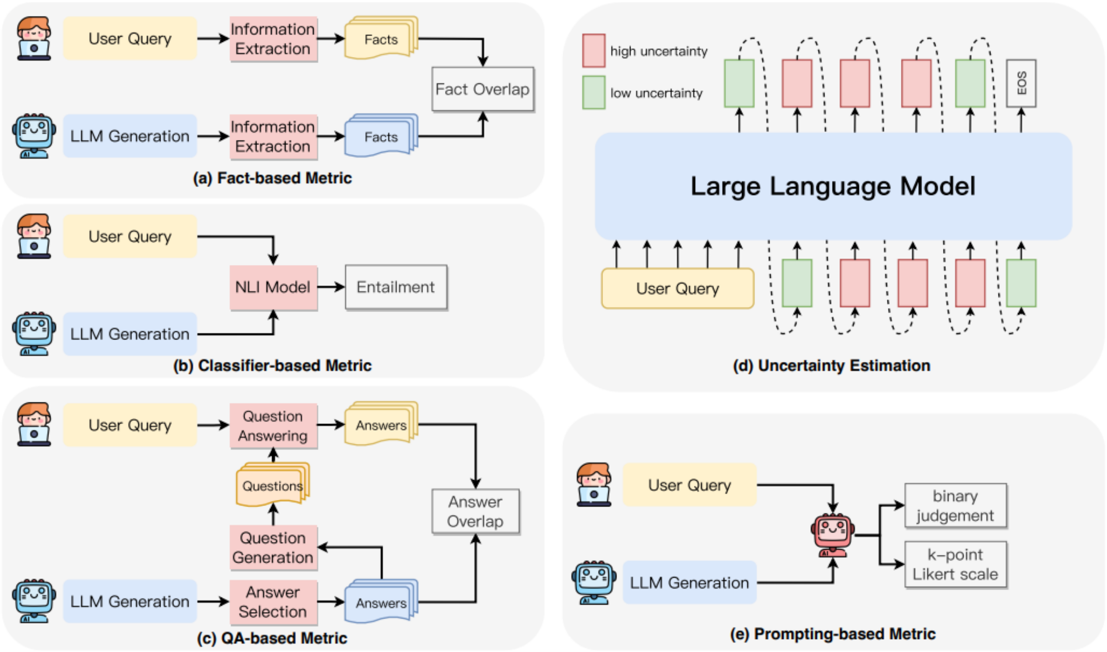


Figure 2. The illustration of detection methods for faithfulness hallucination.

References

[1] HUANG, Lei, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232, 2023.

[2] SUN, Yuhong, et al. Benchmarking Hallucination in Large Language Models based on Unanswerable Math Word Problem. arXiv preprint arXiv:2403.03558, 2024.