CERIAS

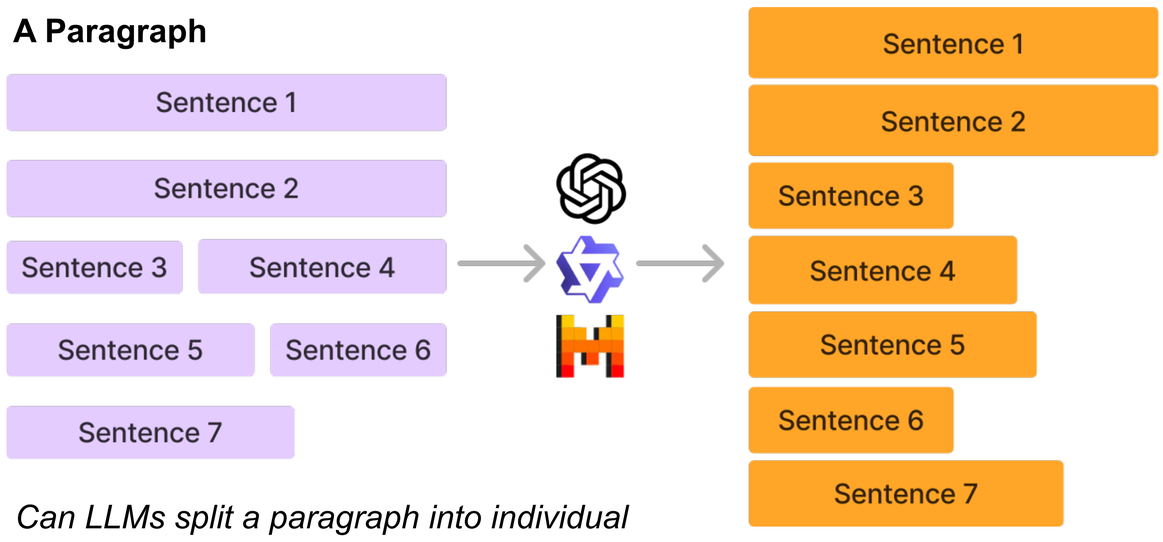
The Center for Education and Research in Information Assurance and Security

Read After Me: Evaluating In-Context Recall Performance of Large Language Models

Yifei Hu (Advisor: Julia Rayz)
Department of Computer and Information Technology, Purdue University

Motivation

- Large Language Models (**LLMs**), such as ChatGPT, have been extensively employed for **in-context question-answering** tasks. While these models are frequently capable of producing plausible responses, they are also susceptible to **hallucination**, which results in factual inaccuracies or trustworthiness concerns with the generated content.
- To perform in-context question-answering effectively, both humans and LLMs should follow a systematic process: **1) identify relevant text from the original document, and subsequently 2) formulate an answer based on the identified text.**
- Although it is challenging to quantitatively assess the quality of LLM-generated answers, we propose a straightforward and efficient benchmark to evaluate the LLMs' capability for in-context recall. Our hypothesis is that **the ability to accurately recall from context correlates positively with the generation of more truthful and less hallucinated answers**



Can LLMs split a paragraph into individual sentences without making mistakes? The task seems very easy, right?

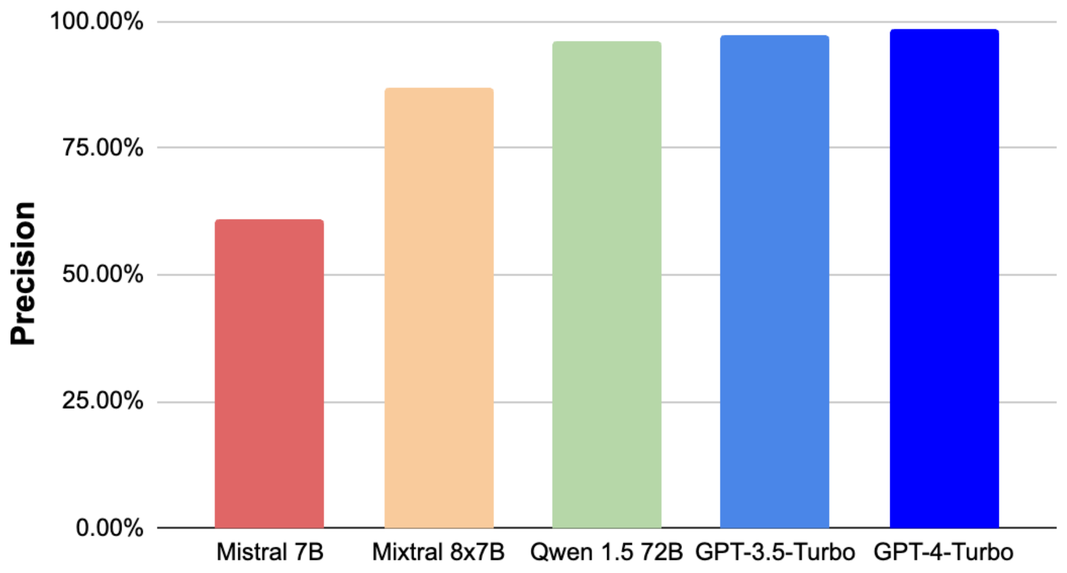
Evaluation Task Design

Unlike the popular **Needle-in-a-haystack** benchmark which inserts irrelevant sentences into a long document and asks the LLMs to perform QA, we propose a more practical setting where all the text in the document is relevant to a certain extent. The task is to ask the LLMs to recall a certain portion of context and split it into individual sentences (which correlates to the first step of performing in-context QA we described in the motivation section). We created the following two sets of tasks:

1. **Sentence Split & Recall (EASY)**: for a given paragraph, split the paragraph into individual sentences
2. **Sentence Split & Recall with Long Context (HARD)**: for a given paragraph in a long document (2500 tokens, 5000 tokens, 8000 token), split the paragraph into individual sentences

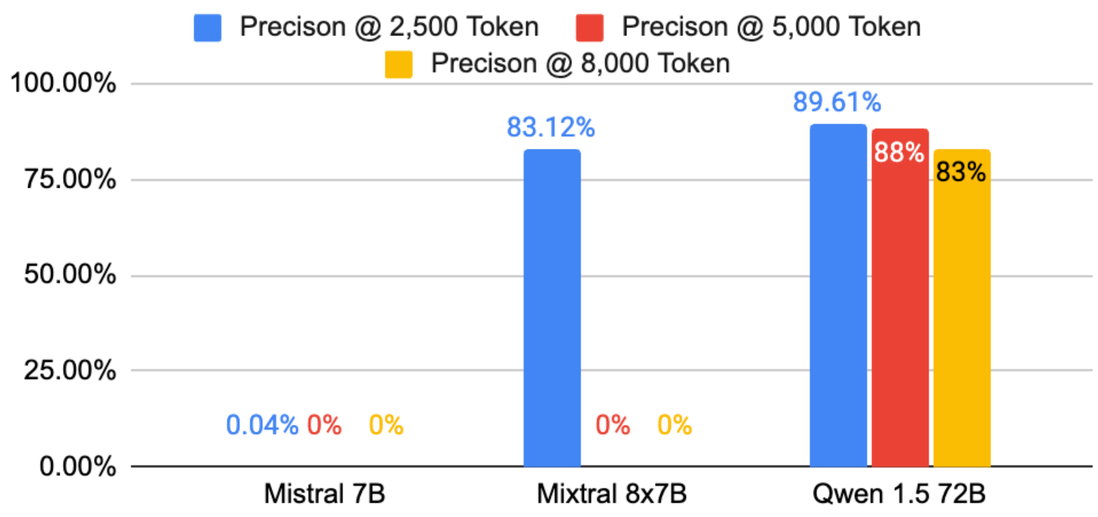
Result on task #1

- The 7B Mistral-Instruct and 8x7B Mixtral-Instruct Models struggled with the simple sentence split & recall task.
- Even the state-of-the-art models like GPT-3.5-Turbo, GPT-4-Turbo, and Qwen-1.5-72B-Chat still failed to perform perfectly.



Result on task #2

- For task #2 Both 7B Mistral-Instruct and 8x7B Mixtral-Instruct were sensitive to longer context and showed significant performance degradation.
- Qwen 72B was more robust to added context yet still experienced slight performance drop.
- Due to the budget limitation, we did not test the proprietary models on task #2.



Findings & Conclusion

- The recalled text can be “mostly” correct but still shows minor errors (paraphrasing, wrong splitting) which are also known as **LLM hallucination**.
- The benchmark tests we proposed are easy to replicate and evaluate the in-context recall capability in a way that is closer to the practical use cases of LLMs.

PURDUE UNIVERSITY

CERIAS