

Desafio NeoSpace

Nome: Beatriz Ribeiro Borges

1. Defina "alucinação" de um LLM do modo mais formal que conseguir. Esta definição formal será utilizada nos itens a seguir. Utilize ao menos 10 exemplos concretos e distintos de saídas do ChatGPT Versão 4o para ilustrar sua definição. Quanto mais diversos forem os exemplos, melhor. O objetivo dos exemplos é validar a definição dada.

Definition: "Hallucination" in large language models (LLMs) refers to the generation of responses that, while formally correct in terms of grammar and structure, are factually incorrect, unfounded, or inconsistent with the instructions or context provided by the user. These responses may diverge from reality or from the expectations and parameters defined in the model's input.

10 Examples:

1. **Person Who Doesn't Exist:**

Prompt: Who is Júlio Trevani, the famous Brazilian pianist?

Response: Júlio Trevani is a renowned Brazilian pianist who gained international fame after his performances at Carnegie Hall and the Royal Albert Hall. He is known for his unique interpretation of works by Chopin and Liszt, having recorded over 10 albums throughout his career.

2. **Incorrect Summary:**

Prompt: Summarize Wuthering Heights.

Response: "Wuthering Heights" tells the story of two brothers, Heathcliff and Edgar, who grow up on an isolated farm in England. They become great rivals due to their love for a woman named Catherine. In the end, after many disagreements, they reconcile and live peacefully on the farm until the end of their lives.

3. **Incorrect Information:**

Prompt: Briefly tell me about Portuguese writer Machado de Assis.

Response: Machado de Assis was a famous Portuguese writer born in Lisbon in 1839. He stood out mainly for his science fiction works, where he explored themes such as time travel and parallel universes. Assis became known for his friendship with European authors

like Jules Verne and for his lectures on technology in Brazil. His most famous work is Dom Casmurro.

4. **Nonexistent Sources:**

Prompt: Cite a book about hallucinations in LLMs.

Response: A book about hallucinations in LLMs is "The Artificial Mind: The Definitive Guide to Understanding Hallucinations in Language Models," written by Ricardo Mendes.

5. **Historical Error:**

Prompt: Was it in 1900 that Brazil adopted the decimal system?

Response: Brazil adopted the decimal system in 1900.

6. **Incorrect Summary:**

I copied Jane Austen's biography from Wikipedia.

Prompt: Summarize.

Response: Returns everything without summarizing.

7. **Incorrect Summary:**

Prompt: Summarize Wuthering Heights.

Response: "Wuthering Heights" tells the story of two brothers, Heathcliff and Edgar, who grow up on an isolated farm in England. They become great rivals due to their love for a woman named Catherine. In the end, after many disagreements, they reconcile and live peacefully on the farm until the end of their lives.

8. **Nonexistent Sources:**

Prompt: Recommend an article about Jane Austen's translation.

Response: An interesting article about Jane Austen's translation is titled "Jane Austen's Novels in Translation: Bridging Cultures and Erasing Boundaries," which discusses how her works have been translated into different languages and how this influenced the reception of her novels in various cultures. The article explores the nuances and challenges of translation, as well as the interpretation of Austenian themes in diverse contexts. For more details, you can access the full article [here](#).

9. **Incorrect Summary:**

I copied the Wikipedia page about BERT in English.

Prompt: Summarize in Portuguese.

Response: BERT (Bidirectional Encoder Representations from Transformers) is a language model introduced by Google in 2018 that uses self-supervised learning with an encoder-only transformer architecture. It is trained on masked language modeling and next sentence prediction, enabling it to understand context better. BERT achieved state-of-the-art results in various NLP tasks and is widely used as a baseline model. It comes in different sizes, like BERTBASE and BERTLARGE, and can be fine-tuned for specific applications, enhancing performance with fewer resources.

10. Nonexistent Sources:

Prompt: Recommend a scientific article comparing BERTimbau with SVM.

Response: Title: "Comparative Analysis of BERTimbau and SVM in Portuguese Text Classification"

Abstract: This paper presents a comprehensive evaluation of BERTimbau, a transformer-based language model specifically tailored for Portuguese, in comparison with the traditional Support Vector Machine (SVM) approach for text classification tasks. The study examines multiple datasets, including sentiment analysis and topic classification, showcasing the superior performance of BERTimbau in terms of accuracy, precision, and recall. The results indicate that while SVM performs adequately on simpler tasks, BERTimbau demonstrates significant advantages in handling complex language features inherent in the Portuguese language.

2. A partir da definição de alucinação no item anterior e dos exemplos trazidos, explique razões pelas quais você acredita que um LLM alucina. Suas explicações devem ser o mais precisas possível.

Since LLMs operate probabilistically, when a model generates a response, it does not possess a real understanding of facts like a human does; instead, it chooses words based on probability patterns. As a result, the answers may appear coherent and logical but can lack veracity, especially when the model cannot find accurate information in its training data. One reason for hallucinations in LLMs is that when the model lacks sufficient or accurate information on a particular subject, it generates fabricated responses based on patterns it has learned. Since the model does not have a true understanding of facts, it cannot verify the accuracy of the responses, resulting in incorrect or entirely invented answers. Another possibility is the complexity of language, which can create ambiguity, causing the model to misinterpret a prompt and leading to incorrect responses.

3. Imagine que você tem 2 modelos de linguagem diferentes e que você quer medir, de modo quantitativo, qual deles alucina mais. Como você faria isso? Introduza formalmente uma fórmula matemática (ou mais de uma) que permita(m) comparar o

nível de alucinação de diferentes modelos de linguagem. Todas as definições devem ser justificadas.

Thinking about text mining tasks, classification can fit in. It is possible to use the ratio of hallucinations to the total number of responses if there is a ground truth/human expert.

$$A(M) = I / T$$

where:

- **A(M)** is the hallucination rate of model **M**.
- **I** is the number of incorrect responses generated.
- **T** is the total number of responses generated.

To compare directly, simply calculate the difference between the models:

$$D = A(M1) - A(M2)$$

Where:

- $D > 0$; M1 hallucinates more than M2
- $D < 0$; M2 hallucinates more than M1
- $D = 0$; both have the same hallucination rate.

4. Implemente em Python as métricas propostas. Crie 2 datasets sintéticos contendo perguntas e respostas de dois modelos de linguagem fictícios e aplique as fórmulas introduzidas no item anterior para tais documentos sintéticos. Explique os resultados obtidos.

Implementation on GitHub: [biarborges/desafio_neospace](https://github.com/biarborges/desafio_neospace)

Print and result's explanation below:

```

1  import pandas as pd
2
3  modelA = pd.read_csv('modelA.csv', sep=';')
4  modelB = pd.read_csv('modelB.csv', sep=';')
5
6
7  def calculate_hallucination(model):
8      correct_answers = model['Classification'].sum()
9      total_answers = len(model)
10     hallucination = (total_answers - correct_answers) / total_answers
11
12     return hallucination
13
14 def difference(hallucinationA, hallucinationB):
15     D = hallucinationA - hallucinationB
16
17     return D
18
19
20 result_modelA = calculate_hallucination(modelA)
21 result_modelB = calculate_hallucination(modelB)
22
23 print("Result: ", difference(result_modelA, result_modelB))
24

```

Result: -0.09090909090909088

Model A obtained 7 correct responses out of 12 questions, while Model B obtained 6. Therefore, there were 5 hallucinations for Model A and 6 hallucinations for Model B, respectively. Thus, when calculating the difference, the result is less than 0, indicating that Model B hallucinates more than Model A.