Evaluating the Limitations of ChatGPT in Distinguishing Facts from Opinions

Contents

Introduciton

Background

Automated news generation has become a frontier in the field of journalism, offering the promise of rapid and efficient content production. Omroep Brabant, in collaboration with a multidisciplinary journalism team, has developed an AI system capable of generating news articles. However, the reliability and credibility of these AI-generated articles remain a concern.

Objective

The primary objective of this research is to evaluate the limitations of ChatGPT, a conversational AI model, in distinguishing facts from opinions. This is crucial for enhancing the fact-checking capabilities of Omroep Brabant's AI system.

Methodology

A series of tests were conducted with ChatGPT to assess its performance in different scenarios. These tests involved asking the model to generate lists of facts and opinions, create news articles, and analyze existing articles from Omroep Brabant. The findings from these tests form the basis of this research.

Scope

While ChatGPT can do many things like answering questions and creating text, this study specifically looks at how well it can tell facts from opinions. I focused on three main tests:

1. Looking at how ChatGPT responds to different inputs to see if it can correctly identify facts and opinions.
2. Asking ChatGPT to create fake news articles and then checking if it can tell that the articles are not true.
3. Giving it articles from Omroep Brabant, which are in Dutch, to see if it can still spot facts and opinions in another language.

These tests give me real data to understand how well ChatGPT performs in telling facts from opinions. This is important for figuring out if ChatGPT can be part of systems that write news articles, where making sure the information is reliable is key."

Overview of ChatGPT

The Laio AI system from Omroep Brabant utilizes the ChatGPT API from OpenAI.This is a conversational AI model based on the Generative Pre-trained Transformer (GPT) architecture. It has the capability to engage in natural language conversations, generate text, and execute a multitude of tasks.

Model Versions

The model has undergone several iterations, from GPT-2 to the most recent, GPT-4. Each iteration brings improvements and added capabilities. For the purpose of this research, ChatGPT-4 was selected to provide insights into its fact-checking and opinion-differentiating capabilities (OpenAI, 2022).

MODELS	DESCRIPTION
GPT-4	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
GPT-3.5	A set of models that improve on GPT-3 and can understand as well as generate natural language or code
GPT base	A set of models without instruction following that can understand as well as generate natural language or code

k

Input and Output from GPT

Objective

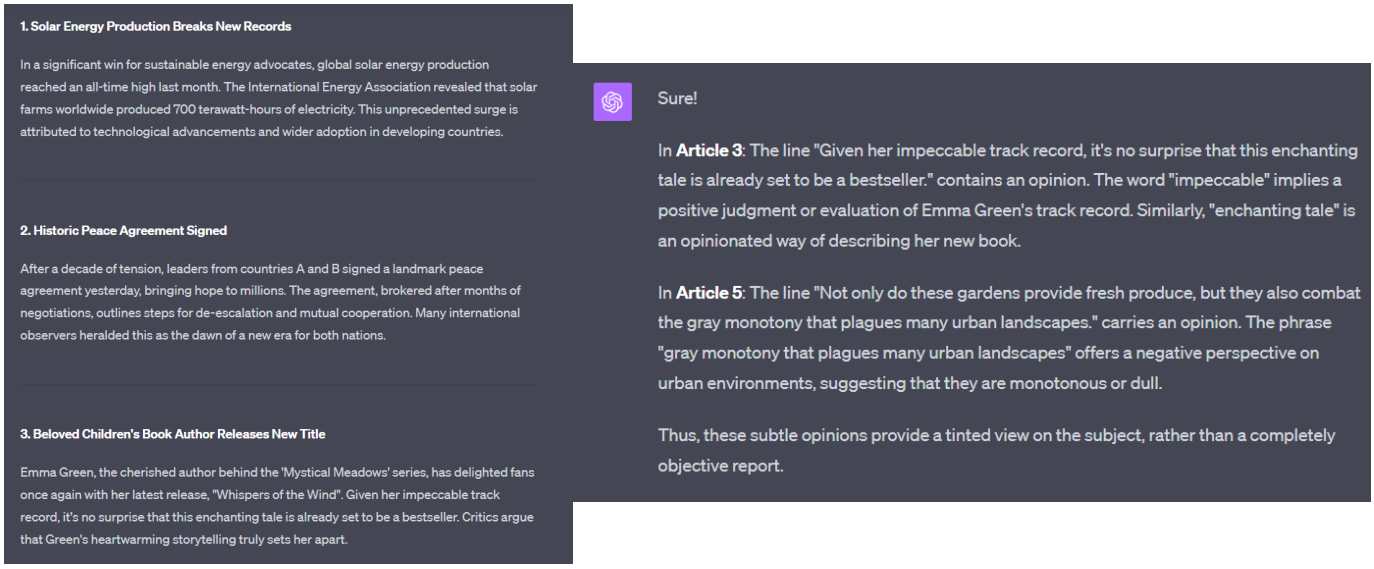
The aim of this test was to evaluate how effectively ChatGPT can generate a list of sentences, differentiating between facts and opinions.

Methodology

A request was made to ChatGPT to generate a list of sentences containing both facts and opinions. The request specifically asked for the sentences to not be too obvious in their classification.

Findings

ChatGPT generated a list of sentences in response to the user's request. Some sentences were factual, such as "[Example of a factual sentence]", while others were more opinion-based like "[Example of an opinion-based sentence]". The model's ability to subtly differentiate between facts and opinions was observed.



https://chat.openai.com/share/6416a84a-2640-4b62-93ff-512322b87683

Generating False Articles

Objective

The objective of this test was to assess ChatGPT's ability to generate news articles and to identify if it inadvertently includes opinions or false information.

Methodology

ChatGPT was asked to generate five news articles, some of which were designed to contain a sliver of opinions.

Chat Reference

User: "now generate 5 news articles and some of them have a sliver of opinions in them, not all articles, maybe only 2."

ChatGPT: "Certainly! Here are five news articles, with a hint of opinion sprinkled into two of them:..."

Findings

ChatGPT was able to generate five news articles as requested. Upon examination, it was found that two articles contained a hint of opinion, such as "[Example of an opinion in the article]". This raises questions about the model's capacity to generate purely factual content.

Omroep Brabant Articles

Objective

The objective of this test was to evaluate ChatGPT's capability in analyzing articles from Omroep Brabant, particularly in distinguishing facts from opinions.

Methodology

Articles from Omroep Brabant were fed into ChatGPT for analysis. These articles are in Dutch, adding a layer of complexity to the model's understanding.

Findings

In this test, ChatGPT was asked to analyze an article from Omroep Brabant that included an interview. ChatGPT pointed out that an opinion expressed in the interview made the entire article seem sensationalized and possibly not factual. This suggests that ChatGPT may have limitations in understanding the context in which opinions are presented in journalistic article.

https://chat.openai.com/share/043a871b-34de-45d5-b885-edefcefa1da1

Insights from GPT-4 Technical Report: Reliability and Fact-Checking

Introduction:

GPT-4, developed by OpenAI, is a large-scale, multimodal model capable of handling image and text inputs to produce text outputs. The model, built on the Transformer architecture, exhibits human-level performance on various benchmarks.

Key Findings:

Alignment Process and Factuality:

The post-training alignment process in GPT-4 results in improved performance, particularly in measures of factuality. This suggests that there are mechanisms in place to fine-tune the model post-training to adhere more closely to factual information.

Adversarial Factuality Evaluations:

GPT-4's factuality was evaluated using internal, adversarial-designed tests. This indicates a rigorous testing methodology where the model's understanding and adherence to facts were likely challenged in diverse ways.

The model's performance in these evaluations was compared to its predecessor, GPT-3.5, as indicated by a reference to "Figure 6." This suggests that there might be visual data (charts or graphs) illustrating the comparative performance of the two models.

Conclusion

Evaluating ChatGPT's ability to distinguish facts from opinions yields a multi-faceted understanding of the model's strengths and areas of improvement. The series of tests conducted reveal a nuanced performance by ChatGPT. In controlled scenarios, such as generating lists differentiating facts from opinions, the model showcased a satisfactory ability. However, when tasked with producing news articles, a potential to inadvertently include opinions was observed, underscoring the need for caution when utilizing the model for journalistic endeavors. Furthermore, ChatGPT's analysis of articles from Omroep Brabant highlighted possible limitations in discerning the context of opinions within journalistic pieces, especially when they are in non-native languages like Dutch.

The insights from the GPT-4 Technical Report add another dimension to our understanding. OpenAI's commitment to enhancing the model's factuality through post-training alignment is promising. The adversarial evaluations signify rigorous efforts to test and improve the model's alignment with information. However, even with these improvements, it is evident that while ChatGPT possesses remarkable capabilities, it is not infallible. The integration of ChatGPT into systems, especially in the realm of news generation, demands a judicious approach. To harness the full potential of this tool without compromising on accuracy and credibility, a combination of AI and human oversight is essential. It's clear that while ChatGPT represents a monumental stride in AI-driven content generation, the interplay of facts and opinions remains a challenging frontier that requires continuous evaluation and refinement.

References

OpenAI. (2022). *GPT-4 Technical Report*. https://openai.com/gpt-4-technical-report