LLM language processing capability analysis

Yanjie Xu
Instructor: Prof. Laura Dietz

1. Introduction

Large language models currently face multiple challenges, such as embedding multiple factual details when generating long answers, which is prone to factual errors, and evaluation methods have difficulty in making precise and accurate assessments of these details. Models frequently produce hallucinations that are inconsistent with real information during the generation process. Most importantly, models are prone to mimic common human misconceptions and generate seemingly reasonable but actually false statements. This not only seriously affects the credibility in practical applications, but also increases the risk of information dissemination. This paper uses a specific task, which is to let the model analyze some user comments on the Internet, obtain a summary text, and then compare it with the standard text to obtain a similarity score to judge the pros and cons of the LLM model. User comments on the Internet are mainly in English, and a few are in other languages. The article also provides a variety of methods to improve the score, including using different models and prompts, and analyzing claims with a mixture of multiple models. We recorded the average and median scores of the models, analyzed the features of claims with a score of 0, and proposed methods to improve the accuracy of the models.

2. Method

This paper primarily employs multiple models and various types of prompts, including the use of Llama-3.3, GPT4, and GTP4O. It also involves model ensembles, such as Llama-3.3+Llama-3.3, GPT4+GPT4, and GPT4O+GTP4O. The prompt types include supporting examples of standardized claims, providing prompts in code mode, emphasizing key points many times, and avoiding negative expressions.

1. Data Types

Reading data files, the data files contain posts, which are comments from users on the Internet, and some are comments from Twitter users. Many of these comments have the characteristics of messy format, lengthy content, unclear topics, and some have some special symbols. The data file also contains a standized claim, which is a summary statement prepared in advance.

2. Prompts Types.

Our method uses multiple prompts,

Explicit instruction prompts
Features: Directly use natural language to clarify instructions, such as asking "tell a joke" or "step-by-step instructions", trying to guide the model to generate specific types of outputs.
Limitations: Instructions may be too simple or vague, and non-professional users often rely too much on intuitive instructions, resulting in unstable output performance in different dialogue situations.

Example dialogue prompts
Features: Embed ideal dialogue examples in prompts, show the expected interaction method through specific demonstrations, and enable the model to imitate this style.
Limitations: Although this method is often more effective in actual results, many non-professional users feel that "borrowing examples" is a bit like "cheating" and are reluctant to adopt this strategy on a large scale.

Repeat prompts
Features: Strengthen the model's memory of specific behaviors by repeating key information or instructions (such as emphasizing a requirement multiple times), so as to expect more consistent output.
Limitations: This strategy has not been fully studied in text generation, and improper use may cause the model output to be lengthy or repetitive.

Templated/code style prompts
Features: Use a format similar to programming or template languages to clearly separate instructions, examples, and other key information, with the aim of improving the structure and consistency of prompts.
Limitations: Although this method helps the model understand instructions more accurately, the style is relatively stiff and may not conform to the natural expression habits of all users.

3. Introduction of LLM

We use Llama-3.3, GPT4, GPT4O, and various combinations.
Llama-3.3: The new 8B and 70B parameter Llama 3 models are a major leap over Llama 2 and establish a new state-of-the-art for LLM models at those scales. Thanks to improvements in pretraining and post-training, Its pretrained and instruction-fine-tuned models are the best models existing today at the 8B and 70B parameter scales. The biggest feature of this model is that there are many free servers that deploy it. And its capabilities are not significantly lower than other paid models. Its model is not weak in scale (70B). The results output by the model also can meet user requirements.

GPT-4: Most influential model released by OpenAI in 2023. We use it because it is a model with great influence in the industry. No other strong competitors were found at the time. It is based on an improved Transformer architecture and uses large-scale pre-training and reinforcement learning techniques. It can process text, images, etc. Although the specific parameter scale and internal details are not disclosed, GPT-4 still has high performance, accuracy, and robustness in handling complex tasks.
GPT4O: Compared with GPT-4, the improvement is in openness and flexibility, but not in absolute improvement in overall performance. The biggest contribution of 4o is that it is released in an open-source way. So researchers can study and check. It is very helpful for the large language model community. Thanks to open source, users can fine-tune and optimize for specific tasks or fields to create models that meet their needs. The cost of 4o is also lower than that of 4.
LLM+LLM：We use a combination of models, where the first model filters out meaningless characters in the text, translates the text (if it is not in English), and organizes the messy text into a paragraph. Then the second model summarizes the paragraph into a standard claim.

4. Experimental Procedure

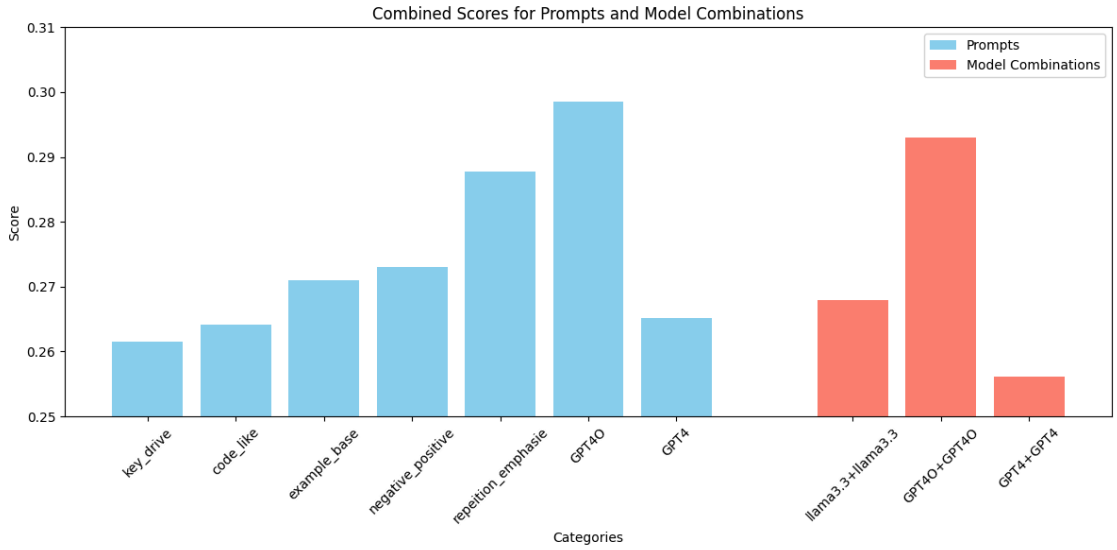
We read the post and standlized claim from the data file line by line. Then we use different APIs to feed the post and prompts to the LLM model. The model returns a normalized claim, and then we compare the normalized claim with the standlized claim. The program used for comparison is METEOR.
METEOR: In 2005, Alon Lavie and Satanjeev Banerjee created METEOR with the goal of surpassing BLEU and ROUGE through the incorporation of synonyms and paraphrases in the assessment process. It is an automatic metric for machine translation evaluation that is based on a generalized concept of unigram matching between machine-produced translation and human-produced reference translations. Unigrams can be matched based on their surface forms, stemmed forms, and meanings; furthermore, METEOR can be easily extended to include more advanced matching strategies. Once all generalized unigram matches between the two strings have been found, METEOR computes a score for this matching using a combination of unigram-precision, unigram-recall, and a measure of fragmentation that is designed to directly capture how well-ordered the matched words in the machine translation are in relation to the reference.

Through METEOR, we can get a score. The higher the score, the smaller the difference between the two claims, which means that the model has a stronger text summarization ability or the prompts are better.
Finally, we calculate the average and median scores of all claims. We found that 4o has the best performance, and Repeat prompts has the best performance in the llama3.3 model. At the same time, we analyzed some claims with a score of 0. We found that these posts have the following characteristics.
1. The overall text is very messy. These posts have many meaningless symbols, just like a cat randomly typing letters on the keyboard.

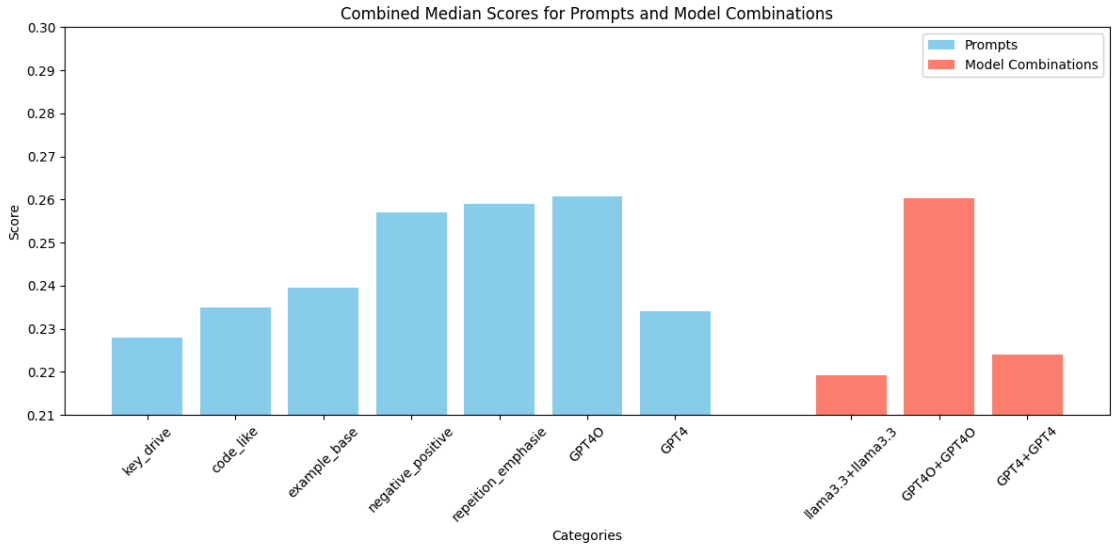
2. The subject is unclear. We found it difficult to find a clear subject.
3. The standlized claims provided in the data file are very different from the post. For example, the post talks about the role of alcohol in fighting the new coronavirus, and the standlized claim says that the “World Health Organization believes that alcohol is very effective in fighting the virus”.

4. experimental results.

Through experiments, we found that the average score is always higher than the median. This means that some claims have achieved very high scores, which raises the overall average score. Both gpt4o and gpt4 use Repeat emphasie prompts. The follow picture shows the overall.



We also calculated the median. The following is the median comparison. We found that the overall score dropped by about 0.03 points.



Zero Score Claims Analyze: We analyzed the characteristics of the claims that scored 0. These posts are extremely chaotic, interspersed with many meaningless symbols that may have been deliberately inserted by users to emphasize certain details. Below is an example of a post intermingled with meaningless symbols.

1. **meaningless symbols**

post:
“Talk the shit @ktthearchdegree rp @nutritionalhealing
Soursop leaves can reverse and prevent dis-ease.

Like, comment & tag a source for Soursop leaves or a beverage provider.

_____
#eatingright #foodismedicine #healthylifestyle #makehealthgreatagain #eatyourmedicine #vegan #veganinatl #vegangang #eatmoregreens #veganfoods #yourhealthisyourwealth #makefoodrealagain #plantbasedatl #hiphopculture #atlantavendors #atlveganfood #veganformyhealth #atlantavegan #atlantavegans #afrovegan #blackandvegan #veganhiphop #makehiphopgreatagain #blackvegansofig #alkalinediet #plantbased #electricfoods #veganhealth #soursop #drsebi Talk the shit @ktthearchdegree rp @nutritionalhealing
Soursop leaves can reverse and prevent dis-ease.”

normalized claim:
Alternative natural treatments to cure serious health conditions

Normalized claim of GPT4:
Soursop leaves can potentially reverse and prevent diseases. Share your sources for Soursop leaves or beverage providers.

From the post, it can be seen that there are many tags following the '#' symbol. After removing the tags, the model's score increased by 0.02, which indicates that the chaotic symbols affected the model's judgment.

2. **Unclear subject:**

Post:
Sonko being Sonko, why do you put hennessy in the donated food pack? Listen #Covid19 Sonko being Sonko, why do you put hennessy in the donated food pack? Listen #Covid19 Sonko being Sonko, why do you put hennessy in the donated food pack? Listen #Covid19 None

normalized claim:
WHO recommends drinking alcohol to prevent coronavirus

normalized claim of GPT4:
'Concerns raised over the inclusion of Hennessy in food packs donated by Sonko during the Covid-19 pandemic.'

Through this example, we can see that the subject in the post is inconsistent with the subject in the normalized claim. The normalized claim generated by GPT-4 appears to be closer to the meaning in the post. The standardized claim provided in the data file seems overly metaphorical, which raises concerns about the correctness of the standardized claims in the data file.

5. Conclusion

Large language models have developed rapidly, but research into the correctness of model outputs has always been a challenge. We analyzed which factors affect the accuracy of model results by having the model summarize posts. The results show that different prompts can influence the scores, and the model's strength also plays a role. However, the combined multiple models is not as good as that of a single model. In the future, more research will focus on the capabilities of the models themselves and the types of prompts. At the same time, we found that the normalized claims in the data file are not entirely correct. In some cases, humans might be more inclined to accept the model's output. We will need a more standardized data file to support further work.

Citation

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. 1Advances in neural information processing systems, 233, 9459-9474.
Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, & Xiangru Tang. (2023). Survey on Factuality in Large Language Models: Methods, Evaluation and Challenges. arXiv preprint arXiv:2311.12526.
Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2019). Evaluating factual consistency of abstractive text summarization. arXiv preprint arXiv:1910.12840.
Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou & Songlin Hu. (2023).Are Large Language Models Good Fact Checkers: A Preliminary Study. arXiv preprint arXiv:2212.10097.