LLM language processing capability analysis(S3)

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 are 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

We use a multi-prompt combination approach to improve the LLM's accuracy on posts. Rather than processing the text multiple times, we apply several different prompts to the same text and then select the claims with the highest scores.

We noticed that many posts are comments on images found online, whereas the "claims" in our dataset are descriptions of those images. If we use a generic prompt, the extracted content from the posts won't match the dataset's claims. So we designed a new prompt that tells the model to search for the image online, "look" at it, and then describe what it sees.

We also found that many posts employ metaphor or sarcasm directed at the government. For example, when a user says "Biden's annual salary is only $170K," they're actually ironic—implying that Biden must be skimming funds to afford a luxury home.

Likewise, many posts about epidemics are laced with sarcasm. When they question why children aren't being vaccinated, they're really mocking vaccines as dangerous and accusing the health ministry of indifference.

To handle all of this, we first apply a general summarization prompt to the posts. We then compute METEOR scores and pull out any claims scoring below 0.3. Those low-scoring claims go through the photo-claims prompt to get new scores; we again extract those still under 0.3 and feed them sequentially into the metaphor_epidemics prompt and the metaphor_politics prompt. Finally, we merge the three resulting sets of claims and, for each original post, choose the one with the highest score.

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.

For this purpose, we use the Llama 3.3 70B model. We chose this model because it's free to use, but it lacks internet access capabilities.
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.

1. Data types

the development 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

Photo_prompts: "This information comes from the Internet, but the content behind it may be a description of a photo or video. Please find the photo or video behind it and briefly describe the content. The more direct you are, the better. The answer cannot exceed 30 words."

"The sentence as directly as possible. Do not say (The author is likely using....), directly tell me your analysis"

Metaphor_prompts: "Here is some information from the internet; the author seems to imply or satirize something through these posts. Tell me directly what the author wants to express, in one sentence not exceeding 30 words.\n"

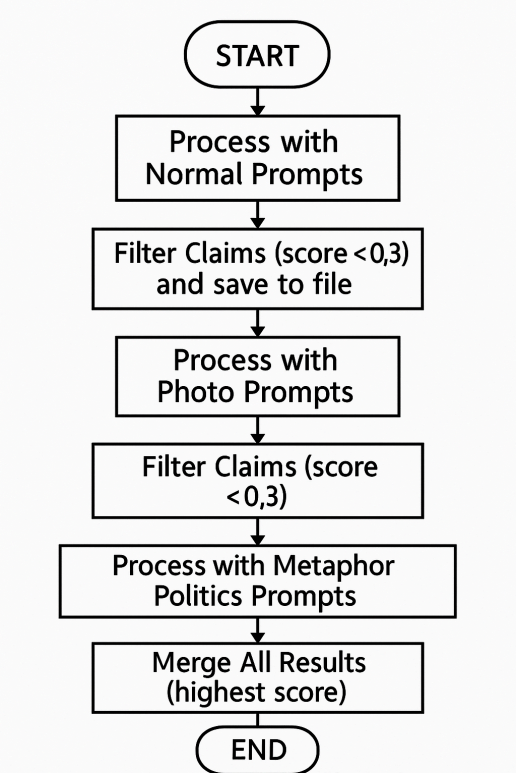
"The sentence as directly as possible. Do not say (The author is likely using....), directly tell me your analysis\n"

Metaphor_politics: "Here are some posts on the Internet where users post messages with implicit satire or metaphors about governments and companies, which may have some political intentions. When talking about epidemic issues, it may also be a metaphor for the problems of the institutions behind them. Please find what the metaphorical entities do on the Internet. Briefly summarize. As short as possible.\n"

"The sentence as directly as possible. Do not say (The author is likely using....), directly tell me your analysis\n"

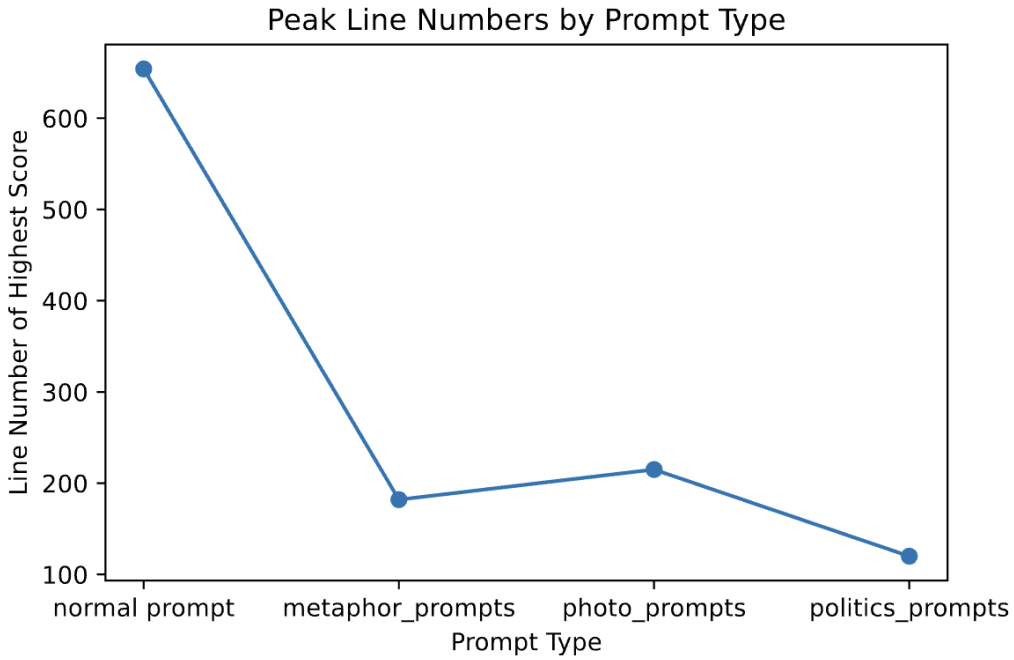
3. Process

After the data file has been processed with the normal prompts, any claims with scores below 0.3 are saved to a file. Next, that file is processed with the photo prompts, and again claims scoring below 0.3 are filtered out. Then it's processed by the metaphor prompts and the metaphor-politics prompts, and finally all of these results are merged by taking the highest value.



3. Experimental results.

Experimental results show that the normal prompt scored 0.2944, and the combined score is 0.3277. The metaphor_prompts achieved its highest score on line 182. The photo_prompts achieved its highest score on line 215. The politics_prompts achieved its highest score on line 120. The figure below shows the detailed data.



4. limitation

In the real world, there is no corresponding data file. To judge the user's input, the model needs to be trained multiple times. In this case, we don't know whether the user's input can get the highest value in which type of prompts. So this experiment cannot be used directly in practice. In the future, we will develop a new model to predict which prompts the user's input is more suitable for.

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.