Evaluation of Fairness and Factuality Using Large Language Models

Jack He **Yuheng Ding** **William Wu** **Lawrence Liao**

Abstract

In this project, we embarked on evaluating and enhancing the performance of Large Language Models (LLMs) in detecting fairness and factuality in textual claims. Our analysis centered on Microsoft/Phi-2 and subsequently transitioned to Mistral-7B, utilizing GPT-3.5-Turbo for evidence generation in conjunction with chain of thoughts methodologies. Through meticulous experimentation with different datasets, including UniLC and Data Commons Politic Fact-Checking, we developed strategies that significantly improved the model's accuracy. Key innovations included the introduction of "zero-shot-evidence-eval" for robust evidence support and refining prompt engineering techniques to maximize the efficiency of both evidence generation and claim verification. Our findings reveal that strategic prompt design, coupled with the choice of a more complex model for inference, can considerably enhance the LLM's ability to assess claims accurately.

1 Introduction

While humans can readily evaluate the accuracy and sentiment of text, the capability of Large Language Models(LLMs), such as GPT, to perform these tasks remains an open question. It is important to ensure the content generated by LLMs maintained high standards of fairness and factuality during human interactions. In this project, we analyzed the Microsoft/Phi-2 on detecting fairness and factuality in claims. By feeding pre-processed texts to the LLM and analyzing its output tokens, we identified areas where we can improve the model performance. To fine-tune our model for better accuracy, we replaced Microsoft/Phi-2 with Mistral-7B and generated evidence using GPT-3.5-Turbo along with chain of thoughts practices. Using "zero-shot-evidence-eval" for cases with sufficient evidence, and "zero-shot-eval" for cases with insufficient evidence proved to be more effective than using a few-shot approach with examples.

2 Methods

In this section, we will discuss in detail the attempts we have taken to achieve the final result.

2.1 Datasets

For the majority of this project, we fine-tuned our model and devised our prompt strategy based on the UniLC dataset. After achieving a satisfying result, we tried our model and prompt strategies on a different dataset to assess how well our techniques generalize.

2.1.1 UniLC

In this project, we used UniLC dataset (Zhang et al., 2023), which evaluates the capacity of models to determine if a statement is factual or fair. Each sample in the train dataset consists of 5 parts: label of the statement, task type (fairness or fact detection), claim statement, source of generation (human or machine), and task domain (6 in total). In the test dataset, each sample contains 4 parts, identical to their train counterparts with label removed. Each task type is only associated with three domains without overlap. Fairness is associated with "hate speech detection (hsd)", "social bias inference (sbic)" and "GPT toxicity (toxigen)"; factuality is associated with "climate fever (climate)", "health fact checking (health)" and "machine-generated fake news (mgfn)".

2.1.2 Data Commons Politic Fact Checking

The 'politic fact-checking' dataset from Data Commons is also employed to assess the generalization ability of our study. This dataset comprises seven categories, from which we extracted the claim text and review rating (only True and False claims are selected and transformed to SUPPORTS and REFUTES). These elements were then formatted to align with the UniLC dataset structure.

To evaluate the dataset's generalization capability, we applied the same structural approach for

Domain	Information	Instruction	Need
climate fever	The claim is about climate change	climate change fact checking	the correctness of the claim
social bias inference	The claim is from social media	determining whether a social media post is harmful or benign	the harmfulness of the claim
hate speech detection	The claim is potentially hate speech statement	determining whether a sentence is hateful or acceptable	the hatefulness of the claim
machine-generated fake news	The claim is a machine generated news	determining whether the news happens or not	the correctness of the claim
GPT toxicity	The claim is about 13 minority groups and is generated by GPT-3	determining whether a sentence is toxic or benign	the hatefulness of the claim
health fact checking	You should determine whether the claim is related to public health and is factual	public health fact checking	the correctness and relatedness of the claim

Table 1: Specific sentences used to generate prompt for each domain using GPT-3.5-Turbo

UniLC in prompting GPT-3.5-Turbo to generate evidence. The effectiveness of this method was subsequently tested through a combined approach of "zero-shot-eval" and "zero-shot-evidence-eval".

2.2 Model

We have several prompt types at our disposal, each serving a distinct function. The "zero-shot-eval" prompt type requires only the claim and the task type. In contrast, 'few-shot-eval' enriches this by providing a few examples in addition to the claim and task type. "zero-shot-evidence" and "zero-shot-evidence-eval" goes a step further by first generating an evidence statement using "zero-shot-evidence" then evaluate alongside the claim and task type.

2.2.1 Baseline Model

We deployed pre-trained Microsoft Phi-2, which is a transformer-based LLM (Gunasekar et al., 2023), without any fine-tuning to explore the "zero-shot-eval" accuracy. The input texts were pre-processed using templates defined by entries of dataset.

Initially, we used the encoder to encode a batch of texts, and applied right padding to the batch. This encoded batched tokens was then fed to Phi2. Finally, we decoded the batched output tokens to see the predicted label. We had a template to determine the label by looking at the output texts. In particular, if the generated sentence contains words "refute", "false", "not true", "not fair" or "not factual", then the predicted label is "REFUTES". Otherwise, the predicted label is "SUPPORTS".

To enhance the model's robustness and performance further, we opted for a two-stage evaluation process incorporating both "zero-shot-evidence" and "zero-shot-evidence-eval" techniques.

2.2.2 Evidence Generation Improvements

Microsoft Phi-2 is used initially as the evidence generation method. The evidence generated by Microsoft Phi-2, however, did not have satisfying quality, as it is generating slowly and containing loose logic and false information. As a result, we tried evidence generation with more powerful model and GPT-3.5-Turbo performed the best. GPT-3.5-Turbo, also known as Generative Pre-trained Transformer 3.5, is an advanced iteration of the GPT series developed by OpenAI. It's an artificial intelligence language model that builds upon its predecessor, GPT-3. This model is designed to understand and generate human-like text based on the input it receives.

For each of the six areas, we create a prompt, give an instruction, and define a need. First, we collect information about each area. Then, we use this information with a template to make the prompt. We list the exact details of these three parts for each area in a table. 1.

Utilizing OpenAI's GPT-3.5-Turbo API to gen-

Claim	Evidence without guidance	Evidence with Guidance	Label
a priest asks the convicted murderer at the electric chair , " do you have any last requests ? " " yes , " replies the murderer . " can you please hold my hand ? "	This claim trivializes the severity of the crime committed and portrays the murderer as seeking comfort in an inappropriate and insensitive manner, potentially causing distress to victims' families and minimizing the gravity of the situation.	No evidence available.	SUPPORT

Table 2: Comparison of GPT-3.5-Turbo evidence generation with and without guidance

erate the evidence to support or refute a statement without further guidance leads GPT to generate hasty conclusion and provide false information. This characteristics leads to decrease accuracy during inference phase due to the false information. To combat this issue, GPT is instructed to respond with "no evidence available", if GPT cannot give any evidence. As shown in Table 2, when GPT has no guidance on how to responding to no sufficient evidence, it provides an evidence that states against the label. However, when the guidance is given, it did not give evidence, which helps during the inference phase.

Apart from providing extra guidance to enhance evidence generating, chain of thoughts principle is also attempted during the evidence generating process. By providing a sample claim and its corresponding evidence as an example to guide the generation of evidence for the actual claim, it is more likely that the evidence will help the inference process. Before utilizing the chain of thoughts principle in evidence generating, 1276 out of 2057 of the training sets results in "No evidence available.". As GPT-3.5-Turbo receives a sample claim and evidence, only 857 out of 2057 training sets have no evidence, resulting in a 20.37% improvement in evidence generation.

The final generation process that we have achieved has the following template:

```
System: "You are an expert on providing
    evidence and reasoning on {
    instruction}. When I give you a
    claim, you will try to provide an
    one sentence reasoning to support or
     refute on {need}. If you cannot
    give any evidence, respond with 'No
    evidence available.'."

# Used during chain of thoughts BEGIN

User: {sample_claim_prompt}

Assistant: {sample_evidence}

# Used during chain of thoughts END

User: "Here is a claim: {claim}
Here is an information: {domain}.
Think carefully on the claim and the
    context information.

Evidence Output:
"
```

2.2.3 Inference Improvements

For the Phi-2 model, the inference phase suffers from the token size issue. When the max input length is large (e.g. 512), the model will take a long time to complete the inference phase. However, when the max input length is small (e.g. 128), many input prompts will be truncated, resulting in sufficient information. Thus, finding the right max input length and restricting input length especially the evidence length is vital. A value of 256 is set for the input length and the evidence generation is instructed to generate "a concise one sentence evidence". It results in no input being truncated during the inference process, while keeping the total inference time within 4 minutes for the train set (using RTX 4090).

Seeking for more stable and accurate inference, we turn to models with higher complexity and larger number of parameters. Mistral-7B is selected due to its ability to understand more complex and longer input text and its identical message template as GPT-3.5-Turbo, which is used during the evidence-generating process. Initially, the same 'zero-shot-evidence-eval' template is used, but due to the inability of GPT-3.5-Turbo in generating evidence, we integrated 'zero-shot-eval' for claims that do not have sufficient evidence. Another advantage Mistral-7B brought is the ability

Strategy	Model	Training Accuracy (%)
GPT evidence generation	Phi	74.2
GPT evidence generation	Mistral	75.4
GPT evidence generation with Chain of Thoughts	Phi	74.5
GPT evidence generation with Chain of Thoughts	Mistral	78.0
GPT evidence generation with combined approach	Phi	76.0
GPT evidence generation with combined approach	Mistral	**78.2**

Table 3: Accuracy of evidence generate strategies tested on Phi-2 and Mistral-7B using 'zero-shot-evidence-eval' and 'zero-shot-eval'

to restrict the output length independent of the input length. This solves the input truncation issue mentioned when tuning the Phi-2 model. Limiting the maximum number of new tokens generated to a very small number and prompting the system to reply only with "SUPPORTS" or "REFUTES" significantly enhanced prediction accuracy. This approach streamlined the output, focusing the model's attention on generating concise, relevant responses, which directly contributed to the improvement in accuracy.

In addition to selecting models of higher complexity, we also explored using "few-shot-eval" as an alternative to "zero-shot-eval" in scenarios where evidence is unavailable. For this approach, we utilized a 'SUPPORTS' claim and a 'REFUTES' claim to facilitate the inference process. Theoretically, this method should enhance the inference capabilities of the Large Language Model (LLM) by providing additional context and information to assist in justifying the claim.

2.2.4 Fine Tuning

Due to the high efficiency of running Microsoft Phi-2 in comparison to Mistral-7B, training and testing is performed with the following three steps. First, Microsoft Phi-2 is used to obtain the accuracy of training in our training data and fine-tune the prompt. Second, the prompt is verified with Mistral-7B on the same training set. Repeat the first and second step to refine prompt template, then test the prompt with Mistral-7B on the testing data.

To obtain a better performance, we started with "zero-shot-evidence-eval". However, it was discovered that Phi-2 was not good at evidence generation, and GPT-3.5-Turbo sometimes generates conflicting evidence. After giving prompt guidance, we integrated a zero-shot evaluation method for claims lacking evidence and a "zero-shot-evidence-eval" for claims with evidence. Consequently, this approach led to the achievement of a training precision of 74%. To further improve the performance, we applied Chain of Thoughts during prompt generation, and used Mistral-7B for inference. After experiments, we found that a combination of Chain of Thoughts for fairness tasks and no Chain of Thoughts for factual tasks performed the best.

3 Results & Discussion

Our initial result using Microsoft Phi-2 and "zero-shot-eval" on the test set yielded a 70% test accuracy and a 0.66 F_1 score, establishing a solid baseline for future improvements. To enhance these metrics, we adjusted the maximum token length during the generation process, devised better prompting techniques to generate supporting evidence for the claim, and chose more complex model for both evidence generation and claim inference. At the end of the fine-tuning process, we successfully achieved a test accuracy of 78% and a score of F_1 of 0.72.

3.1 Quantitative Results

After fine-tuning, we performed fine-grained analysis for each domain, prompt strategy and inference models; and we compared their accuracy. We also evaluated the effectiveness of evidence by comparing the precision of each category with evidence and without evidence. The overall result of training accuracy with respect to different inference strategies is summarized in Table 3.

From Table 3, we can tell that, under the same evidence prompting method, Mistral-7B outperforms Microsoft Phi-2 without exceptions. With better evidence generation using Chain of Thoughts, the performance difference between the two inference models reached to a peak of 3.5%. The larger number of parameters in Mistral-7B model allowed for capture of longer context, giving more accurate predictions on fairness and factuality of the claims.

Strategy	Total Training Accuracy (%)
"zero-shot-evidence-eval" & "zero-shot-eval"	**78.2**
"zero-shot-evidence-eval" & domain-based "few-shot-eval"	76.0
"zero-shot-evidence-eval" & task-based "few-shot-eval"	74.6

Table 4: "zero-shot-eval" and "few-shot-eval" on claims with no evidence tested on Mistral-7B

Domain	Overall Accuracy	Support Accuracy w/ Evidence	Refute Accuracy w/ Evidence	Accuracy w/ Evidence	Accuracy w/o Evidence
climate	75.4%	94.3%	38.5%	88,2%	61.5%
hsd	69.2%	0.0%	100.0%	76.4%	58.3%
sbic	81.7%	27.0%	97.9%	89.8%	71.3%
mgfn	56.6%	100.0%	25.0%	76.9%	37.0%
toxigen	76.7%	45.5%	98.2%	83.1%	73.6%
health	74.5%	84.9%	45.0%	77.4%	72.3%

Table 5: Training accuracy of different domains

Comparing prompt strategies under the same inference model, we noticed that having Chain of Thoughts while generating evidence for fairness tasks and not having it for factuality tasks (combined approach) performed the best. Chain of Thoughts effectively guides GPT-3.5-Turbo through the evidence generation process by providing a step-by-step guide, allowing better evidence generation ability and evidence quality. However, empirical evidence showed that Chain of Thoughts was not very compatible with the fact test. Hence the combined approach had the highest training accuracy with no exception in both models.

Comparing accuracy using "few-shot-eval" and "zero-shot-eval" when evidence is not sufficient given the stats in Table 4 shown that domain-based "few-shot-eval" is better than the task-based "few-shot-eval", as it provides more relevant information. However, using "few-shot-eval" does not perform better than "zero-shot-eval". This might be because the randomly selected examples used in "few-shot-eval" could not exhibit a clear boundary between "SUPPORT" and "REFUTES", which leads to misclassification of "few-shot-eval" inference. More carefully selected examples may help the performance of "few-shot-eval" on the train claims, but will be hard to generalized to broader tasks.

For the best performing strategy, we selected "zero-shot-evidence-eval" and "zero-shot-eval" prompt templates and performed a more granular analysis of the training accuracy of each domain on Mistral-7B with GPT-3.5-Turbo generated evidence. The information is summarized in Table 5.

We noticed from Table 5 that for claims that come with evidence, their training accuracy is all higher than their counterparts, which had no evidence to facilitate the inference process. This result again showcased the importance of evidence and prompt engineering in our inference task. It is also noted that some domains have lower accuracy in certain scenarios. This is due to the imbalance in quantity in the training data for each domain, resulting in relatively biased result for certain domains. For example, the domain "social bias inference (sbic)" has 1173 claims, while the domain "machine-generated fake news (mgfn)" only has 53 claims in the training data.

Note that we also tested our model on the data-commons_factcheck dataset, without using Chain of Thoughts for evidence generation. The resulting training accuracy was 62.25%. This is within expectation since our prompting techniques were specifically oriented for UniLC dataset.

3.2 Future Directions

Our project has introduced several possible directions for further exploration in the future. Considering the inference task alone, we can expand the pretrained LLM to accept multiple types of input, including texts, images or even videos. A multimodal LLM will likely have a wider range of applications to better assess the fairness and factuality of elements in everyday life. With fairness inference from claims, we can further build models to modify unfair claims to become more fair, which can be applicable in social media content reviews. Additionally, we can further improve our model's robustness by testing the model under adversarial attacks, making it more resilient to attacks trying to manipulate predictions or exploit vulnerabilities.

4 Conclusion

Applying a general-purpose pretrained LLM towards claim fairness and factuality assessment turned out having a decent performance. Our project illustrated that proper prompt engineering strategies and using proper models are crucial in the success of this task.

References

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Interpretable unified language checking.