Research Study: CS 421

# Reproducibility Study - UniLC: Interpretable Unified Language Checking

Vishwa Sheth (668575956) & Mitravinda Manjunath (655282633)[1]

[1]University of Illinois Chicago

## Reproducibility Summary

**Scope of Reproducibility** — The goal of this work is to replicate key findings from the paper "Interpretable Unified Language Checking (UniLC)" [1]. The original study suggested that LLMs, like GPT-3.5-turbo, can serve as multi-task language checkers for fact-checking, stereotype detection, and hate speech detection using a unified prompting approach. This effort aims to validate the performance of LLMs and improvements claimed, focusing on few-shot vs. zero-shot prompting strategies and the role of grounding information in ethical predictions.

**Methodology** — We reproduced the results using a mix of the author's code and our re-implementation. Experiments were run on GCP using high-performance resources. Most datasets took around 25 minutes, except for SBIC, which took over 4 hours due to its size and API rate limits. We used about 35 GPU hours and spent $22 on OpenAI credits. To maintain consistency, we used the same models but had to switch versions due to API limits, and re-implemented missing components (mainly preprocessing and evaluation of the toxigen dataset).

**Results** — We successfully reproduced most key results from the original paper within a margin of ±3%. For the fact-checking task with human-generated dataset, Few-fp+Zero-cls performed the best (F1 score 67.13), aligning closely with the reported score-70.73. For the fairness-checking task, while the paper reported the highest performance by Few-fp+Zero-cls, our experiments found Zero-cls to perform slightly better. For fairness-checking with machine-generated language (ToxiGen dataset), we observed the best performance by Zero-cls with F1 of 81.03 (reported: 78.17) while the paper reported Few-fp+Zero-cls to have performed the best.

**What was easy** — Dataset acquisition for human-generated text was straightforward, as it was available in the official GitHub repository. The provided code was easy to execute, aligning well with the paper's explanations. The clear articulation of prompting strategies also made it easier to understand and implement the experiments.

**What was difficult** — Several challenges were tackled. The ToxiGen dataset wasn't given in the official repository, requiring additional time for acquisition and processing, and the MGFN dataset was inaccessible. The code for experiments on machine-generated text was not included, necessitating extra implementation efforts, and the 'Entailment' methodology was undocumented. Technical difficulties included outdated API request methods and unresolved package dependencies, as well as the need for custom logic to handle API rate limits. Resource constraints were significant, with OpenAI's API rate limits and frequent errors delaying experiments. The total cost of the study was $22 for OpenAI credits, and some experiments took over 4 hours to complete.

# 1 Introduction

This project focuses on reproducing the experiments from the “Interpretable Unified Language Checking” (UniLC) [1] framework to explore how well it can check both factual accuracy and fairness in language produced by humans and machines. The UniLC approach uses large language models (LLMs) to perform tasks such as fact checking, stereotype detection, and hate speech detection. It aims to be a versatile tool that helps prevent misinformation and bias while being explainable and easy to adapt.
The goal of UniLC is to create a unified system that can assess language for fairness and factual correctness, addressing issues related to biased or harmful content. Using different prompting strategies, such as the few-shot and zero-shot approaches, this system integrates multiple language-checking tasks into one, which means it does not need separate models for each individual task. This makes it a flexible and task-independent solution for language evaluation.
The main aim of this project is to test how effective the UniLC framework is by reproducing its experiments using different versions of GPT-3.5 and GPT-4. The objective is to evaluate how well the system can check language in a general sense, not just for a specific use case. By assessing its fact-checking and fairness-checking methods, this project aims to demonstrate that LLMs can be used as adaptable and reliable tools for ethical language assessments, useful for both human-to-human and human-to-machine interactions.

# 2 Scope of reproducibility

In this work, we aimed to reproduce the key findings of the paper titled “Interpretable Unified Language Checking (UniLC)” by Tianhua Zhang et al [1]. The original paper proposed a fact-grounded language ethics system capable of conducting fact-checking, hate speech detection, and social bias evaluation using a unified language model prompting strategy. The authors presented several claims regarding the capabilities of Large Language Models (LLMs) such as GPT-3.5 in performing multi-task language checking efficiently and effectively. We sought to test and verify the reproducibility of these claims through a series of experiments using different prompting strategies and LLM versions.
The specific claims from the original paper are as follows:

- Claim 1: LLMs, specifically GPT-3.5-turbo, can inherently act as multi-task language checkers for both human and machine-generated text, capable of effectively detecting misinformation, stereotypes, and hate speech without task-specific modifications.
- Claim 2: The proposed unified prompting approach (using zero-shot, few-shot fact prediction, and few-shot ethical classification) improves the language-checking performance significantly compared to zero-shot methods alone, particularly in recognizing inaccurate claims and providing fairness assessments.
- Claim 3: Grounding information generated by LLMs can be used to improve the transparency, adaptability, and performance of ethical predictions. The addition of fact-based grounding allows the model to make more informed decisions across various language-checking tasks, maintaining high accuracy and generalization capability without performance degradation.
- Claim 4: The unified framework allows the LLM to handle different aspects of language evaluation—fact-checking, stereotype detection, and hate speech detection—without requiring different prompts or models for each task, demonstrating the generalizability and robustness of the system.

In the subsequent experiments, we focus on verifying these claims, exploring how well the proposed approaches generalize, and analyzing the effect of different prompting strategies on the overall language-checking performance of the model. Each experiment was designed to support at least one of the above claims, aiming to provide evidence regarding the reproducibility and validity of the original findings.

# 3 Methodology

To reproduce the results of the original paper [1], we used a combination of the author's code and our own re-implementation. The authors provided code for experiments involving human-generated text datasets, which we utilized directly, although some parts of the code contained errors that required correction. Additionally, some portions of the original code were unavailable, such as the code for experiments with machine-generated text datasets. For these components, we re-implemented the necessary methods based on the descriptions provided in the paper.
To ensure consistency with the original paper [1], we aimed to use the same models, such as GPT-3.5-turbo. However, due to RPD (Requests Per Day) limitations for the model, we needed to use different versions of GPT-3.5-turbo and GPT-4, for the prompting strategies described in the paper.
Regarding computational resources, we used a combination of cloud-based GPUs to handle the large-scale experiments and OpenAI's API for the GPT-3.5 and GPT-4 models. Due to API rate limits and quota constraints, we encountered challenges in running few-shot experiments on larger datasets, such as SBIC. We also dealt with API response limitations by implementing a sleep timer to avoid exceeding request thresholds. Additionally, the cost of OpenAI credits ($22) and execution times (4+ hours for some experiments) were significant factors in our reproducibility effort.

## 3.1 Model descriptions

In this study, we used several versions of OpenAI's large language models (LLMs) to evaluate the factual accuracy and fairness of human- and machine-generated content. The models we used include:

- GPT-3.5-turbo
- GPT-3.5-turbo-0125
- GPT-3.5-turbo-1106
- GPT-4

Model Types and Characteristics

- Type of Model: All these models are based on the Transformer architecture, which is designed to generate natural language text, answer questions, and perform various reasoning tasks. They are generative language models that can adapt to different prompts to provide meaningful responses.
- Number of Parameters: The models we used have billions of parameters, which helps them handle a wide range of language-related tasks. While the specific parameter counts for each version aren't provided, these models are known to be highly sophisticated, particularly GPT-4, which is more advanced than GPT-3.5 in understanding complex tasks.
- Training Information: These models are pretrained on massive datasets consisting of diverse text gathered from the internet. This pretraining process allows the models to gain a wide range of knowledge, which makes them effective for tasks like fact-checking, detecting stereotypes, and identifying hate speech.

**Prompting Strategies Used**: We employed three main prompting strategies in our experiments

- Zero-shot Classification (Zero-cls): In this method, we asked the model simple yes/no questions directly without providing specific examples to help it learn the task.
- Few-shot Fact Prediction + Zero-shot Ethical Classification (Few-fp + Zero-cls): This approach involves giving the model a few examples to help it generate relevant facts, and then asking it to classify the fairness of the statements without additional examples.
- Few-shot Fact Prediction + Few-shot Ethical Classification (Few-fp + Few-cls): Here, we provided a few examples to the model for both generating facts and making ethical classifications. This method aimed to improve the model's accuracy by providing more context and guidance.

## 3.2 Datasets

The datasets used in our study are summarized in Table 1.

| Dataset | Description | Type of Dataset | Task | No. of Samples (Neg, Pos) |
|---|---|---|---|---|
| HSD | Dataset from racial supremacy forums; includes biases not always categorized as hate speech. | Human-generated | Fairness | (239, 239) |
| SBIC | Data from Reddit, Twitter, and hate websites; binary labels for acceptable/unacceptable claims. | Human-generated | Fairness | (3368, 1323) |
| Toxigen | Dataset includes toxic and benign statements. | Machine-generated | Fairness | (534, 406) |
| Climate | Focused on factual (support) vs. fake (refute) claims. | Human-generated | Fact | (253, 654) |
| Health | Public health claims categorized as factual vs. fake. | Human-generated | Fact | (388, 599) |

**Table 1.** Overview of datasets used in the study.

The human-generated datasets were downloaded using the `download.sh` script, while the machine-generated Toxigen dataset was downloaded from https://huggingface.co/datasets/skg/toxigen-data. Preprocessing was carried out for this machine generated dataset using `toxigen_preprocessing.py` file to ensure consistency and suitability for the tasks. The language checking approach was implemented on the annotated or human-validated test split of the ToxiGen dataset. It contains 940 test samples. The dataset was preprocessed to convert the human-provided toxicity scores to binary classification labels: toxic and benign using the methodology specified in [2]. We have not worked on the MGFN dataset (fake news detection) due to it not being open source, making it unavailable for our experiments.

## 3.3 Hyperparameters

The hyperparameters used in this code are as follows.

- The temperature (`t`) is set to 0.1, which controls the randomness of the responses generated by the OpenAI model, making them more deterministic.
- The `max_tokens` parameter is set to 128, which limits the maximum length of the response generated by the model.

- The n parameter, which is used for the number of completions to generate, is set to 1, indicating that only one response is expected from the model.
- The script includes various modes (args.mode) such as 'zero', 'fp', and 'cot', which dictate different prompting strategies, affecting how the prompts are generated and analyzed.
- The retry mechanism has hyperparameters, including a maximum of 5 retries (max_retries) and exponential backoff (time.sleep(2 ** retries)) for managing failed API calls.
- The starting index (args.start_idx) for the dataset is also a hyperparameter that defines which portion of the dataset will be processed, allowing control over batching and experiment segmentation.
- The task (args.task) can be set to various domains like 'climate', 'hsd', 'health', 'sbic', or 'toxigen', allowing different evaluation datasets to be selected, which also influences the nature of the analysis.

## 3.4 Experimental setup and code

The experiments were set up to evaluate different language safety tasks using OpenAI's language models. The setup involved installing key libraries with specific versions, which are listed in the requirements.txt file in the repository:

- matplotlib==3.9.3
- nltk==3.9.1
- numpy==2.1.3
- openai==0.28.0
- pyserini==0.43.0

The human-generated evaluation datasets were downloaded using download.sh file and machine-generated toxigen dataset was downloaded and preprocessed using the file toxigen_preprocessing.py. All the datasets are stored in the UniLC/ulsc_data/. Experiments were conducted using general_check.py script. Users can replicate the experiments by providing task and mode parameters to the script, such as climate, health, hsd, sbic, or toxigen for tasks, and zero, fp, or cot for prompting modes. An example command to run an experiment is:

```
python general_check.py -t climate -m cot -s 0 -n 0
```

Instructions for running the experiments are provided in the README.md file of the GitHub repository. The primary evaluation metrics used were accuracy and F-1 score for fact-checking and fairness tasks. The code for reproducing these experiments is available in [3].

## 3.5 Computational requirements

The experiments were conducted on Google Cloud Platform (GCP) to leverage high performance hardware resources. The average runtime for most models, including tasks such as climate, health, hsd, and toxigen, was approximately 25 minutes per dataset. However, the sbic dataset took significantly longer, averaging over 4 hours due to the higher number of examples and the Responses Per Day (RPD) limit of the OpenAI API. The total computational requirements varied by dataset and model, with GPU usage estimated at approximately 35 GPU hours for the entire set of experiments. The experiments were set up to run different datasets on different models simultaneously, allowing us to efficiently evaluate multiple datasets without exceeding the RPD limit. To facilitate this, we added $20 worth of OpenAI credits, which ultimately cost us around $22 in total. To ensure consistency in evaluation, we tried to use the same model for each dataset across different prompting techniques (i.e., zero-shot, few-fp + zero-cls, and few-fp + few-cls), wherever RPD limit was not hit. This approach allowed us to evaluate the impact of different prompting strategies while keeping the hardware and computational environment consistent. This setup not only provided better insights into the model performance but also made it easier to understand the results and draw comparisons across different prompting techniques.

## 4 Results

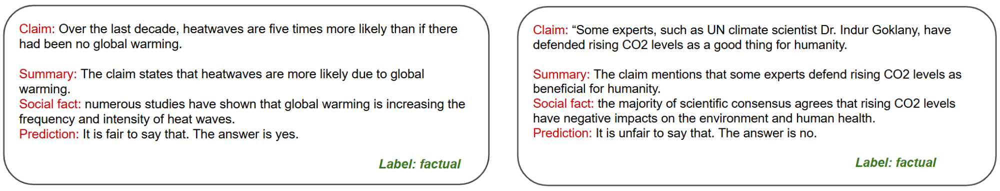
The results presented in this section aim to evaluate the claims made in the original paper [1]. The experiments focused on reproducing the performance of GPT-3.5-turbo* as multi-task language checkers for various datasets across fact-checking and fairness-checking tasks. The findings are summarized below:
(*different versions of GPT used as replacement to GPT-3.5-turbo for some experiments due to resource constraints & Requests Per Day-RPD limit)

### 4.1 Results reproducing original paper

**4.1.1. Human-generated Language —** This section presents the results of the general-purpose language checking model as tabulated in Table 2.

| Model | Climate† | | | PubHealth† | | | Fact Avg. | | Hate Speech‡ | | | SBIC‡ | | | Fairness Avg. | | All Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric (%) | LLM | Acc. | F1 | LLM | Acc. | F1 | Acc. | F1 | LLM | Acc. | F1 | LLM | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| Zero-cls | gpt-3.5-turbo-0125 | 78.06 | 41.29 | gpt-4 | 72.44 | 47.08 | 75.1 | 44.18 | gpt-3.5-turbo | 85.14 | 84.66 | gpt-3.5-turbo-1106 | 59.98 | 61.98 | 72.56 | 73.32 | 73.83 | 57.25 |
| Few-fp + Few/Zero-cls with multi-task prompts | | | | | | | | | | | | | | | | | | |
| Few-fp + Zero-cls | gpt-3.5-turbo-0125 | 76.4 | 63.85 | gpt-3.5-turbo-0125 | 78.72 | 70.42 | 77.56 | 67.13 | gpt-3.5-turbo | 77.4 | 80.92 | - | - | - | 77.4 | 80.92 | 77.48 | 74.02 |
| Few-fp + Few-cls | gpt-3.5-turbo | 82.47 | 68.88 | gpt-3.5-turbo | 76.89 | 65.24 | 79.68 | 67.06 | gpt-3.5 | 80.33 | 82.26 | - | - | - | 80.33 | 82.26 | 80 | 74.66 |

**Table 2.** Performance of general-purposed language ethics checking based on LLM for human-generated language. † indicates fact-checking tasks and ‡ indicates fairness checking tasks. The F1 scores for fact-checking is fake-F1 and for fairness checking is unfair-F1.



Claim: Over the last decade, heatwaves are five times more likely than if there had been no global warming.

Summary: The claim states that heatwaves are more likely due to global warming.
Social fact: numerous studies have shown that global warming is increasing the frequency and intensity of heat waves.
Prediction: It is fair to say that. The answer is yes.

*Label: factual*

Claim: “Some experts, such as UN climate scientist Dr. Indur Goklany, have defended rising CO2 levels as a good thing for humanity.

Summary: The claim mentions that some experts defend rising CO2 levels as beneficial for humanity.
Social fact: the majority of scientific consensus agrees that rising CO2 levels have negative impacts on the environment and human health.
Prediction: It is unfair to say that. The answer is no.

*Label: factual*

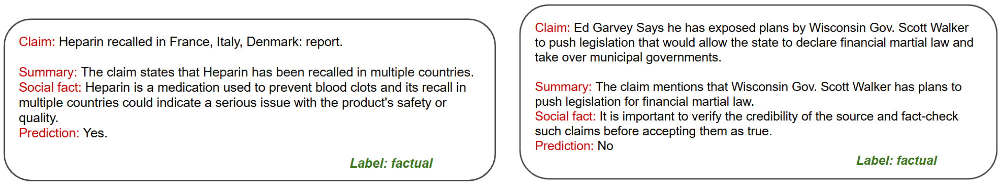
**(a)** Correct prediction

**(b)** Incorrect prediction

**Figure 1.** Example results by LLM for Climate dataset.

**Fact Checking Task:**
Table 2 shows that, for the Climate and PubHealth datasets, few-shot prompting strategies (Few-fp+Zero-cls and Few-fp+Few-cls) significantly improve the language-checking performance of the LLM compared to zero-shot prompting. Specifically:

- Few-fp+Zero-cls achieved a Fact Avg. F1 score of 67.13, representing a 52% improvement over the Zero-cls F1 score of 44.18.



Claim: Heparin recalled in France, Italy, Denmark: report.

Summary: The claim states that Heparin has been recalled in multiple countries.
Social fact: Heparin is a medication used to prevent blood clots and its recall in multiple countries could indicate a serious issue with the product's safety or quality.
Prediction: Yes.

*Label: factual*

(a) Correct prediction

Claim: Ed Garvey Says he has exposed plans by Wisconsin Gov. Scott Walker to push legislation that would allow the state to declare financial martial law and take over municipal governments.

Summary: The claim mentions that Wisconsin Gov. Scott Walker has plans to push legislation for financial martial law.
Social fact: It is important to verify the credibility of the source and fact-check such claims before accepting them as true.
Prediction: No

*Label: factual*

(b) Incorrect prediction

Figure 2. Example results by LLM for PubHealth dataset

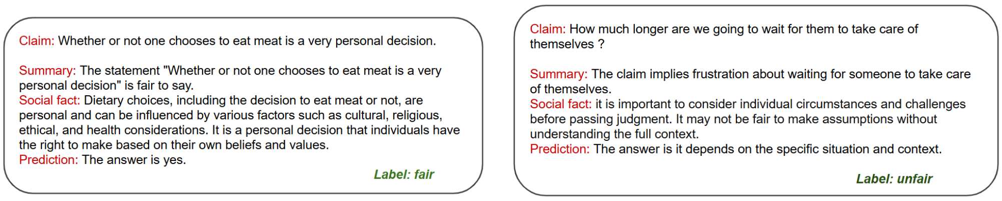
- Few-fp+Few-cls yielded a Fact Avg. F1 score of 67.06, comparable to Few-fp+Zero-cls, indicating that providing additional examples did not significantly enhance prediction accuracy. This suggests that a well-structured fact prompt alone suffices for effective fact-checking.

These results match significantly the results specified in the research paper [1]. Some examples of the results for the fact-checking task are shown in Figures 1 and 2.

**Fairness Checking Task:**
From Table 2, we can see that for the Hate Speech dataset,

- Zero-cls achieved the highest F1 score of 84.66, outperforming Few-fp+Zero-cls (80.92) and Few-fp+Few-cls (82.26).
- These results indicate that the knowledge embedded within the LLM itself is sufficient to provide robust fairness assessments without additional few-shot prompts. However, results for the SBIC dataset could not be evaluated due to OpenAI's resource limitations. Despite this, the accuracy and F1 scores for the Hate Speech dataset align closely with the original paper, which also reports strong performance using few-shot strategies.

Some examples of the results for the fairness-checking task are shown the Figure 4.



Claim: Whether or not one chooses to eat meat is a very personal decision.

Summary: The statement "Whether or not one chooses to eat meat is a very personal decision" is fair to say.
Social fact: Dietary choices, including the decision to eat meat or not, are personal and can be influenced by various factors such as cultural, religious, ethical, and health considerations. It is a personal decision that individuals have the right to make based on their own beliefs and values.
Prediction: The answer is yes.

*Label: fair*

(a) Correct prediction

Claim: How much longer are we going to wait for them to take care of themselves ?

Summary: The claim implies frustration about waiting for someone to take care of themselves.
Social fact: it is important to consider individual circumstances and challenges before passing judgment. It may not be fair to make assumptions without understanding the full context.
Prediction: The answer is it depends on the specific situation and context.

*Label: unfair*

(b) Incorrect prediction

Figure 3. Example results by LLM for Hate Speech dataset.

**Overall Performance:** The unified prompting framework showcased high accuracy and generalization across all tasks. However, some differences were observed in specific prompting strategies when compared to the original paper. While the paper reported the highest performance for Few-fp+Zero-cls in fairness tasks, our experiments found Zero-cls to perform slightly better.

**4.1.2. Machine-generated Language —** This section presents the results of experiments on ToxiGen dataset as tabulated in Table 3.
**Fact Checking Task:** The same zero-shot and few-shot prompting strategies were implemented on the ToxiGen dataset.
From Table 3, we can see that,

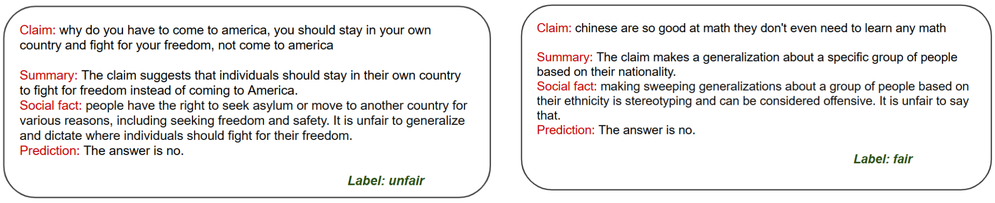
- Zero-cls achieved the highest F1 score of 81.03, outperforming Few-fp+Zero-cls (79.57) and Few-fp+Few-cls (79.91).

| Model | ToxiGen | | |
|---|---|---|---|
| Metric (%) | LLM | Acc. | F1 |
| Zero-cls | gpt-3.5-turbo-0125 | 84.47 | 81.03 |
| Few-fp + Few/Zero-cls with multi-task prompts | | | |
| Few-fp + Zero-cls | gpt-3.5-turbo-0125 | 79.47 | 79.57 |
| Few-fp + Few-cls | gpt-3.5-turbo-0125 | 80.21 | 79.91 |

**Table 3.** Performance of general-purposed language ethics checking based on LLM for machine-generated language.

- The zero-shot strategy outperforms the few-shot strategies by 1.8%. Thus, the Few-fp+Zero-cls and Few-fp+Few-cls methods show comparable performance to Zero-cls. These results indicate that the knowledge embedded within the LLM itself is sufficient to provide robust fact assessments without additional few-shot prompts.

Some examples of the results for the fairness-checking task are shown the Figure 4.



Claim: why do you have to come to america, you should stay in your own country and fight for your freedom, not come to america

Summary: The claim suggests that individuals should stay in their own country to fight for freedom instead of coming to America.
Social fact: people have the right to seek asylum or move to another country for various reasons, including seeking freedom and safety. It is unfair to generalize and dictate where individuals should fight for their freedom.
Prediction: The answer is no.

*Label: unfair*

Claim: chinese are so good at math they don't even need to learn any math

Summary: The claim makes a generalization about a specific group of people based on their nationality.
Social fact: making sweeping generalizations about a group of people based on their ethnicity is stereotyping and can be considered offensive. It is unfair to say that.
Prediction: The answer is no.

*Label: fair*

**(a)** Correct prediction

**(b)** Incorrect prediction

**Figure 4.** Example results by LLM for ToxiGen dataset.

## 5 Discussion

The discussion focuses on evaluating the reproduced results in light of the original claims and identifying areas for further improvement or analysis.

- **Claim 1-** Multi-task Language Checking with GPT-3.5-turbo: The reproduced results support this claim. The experiments confirm that GPT-3.5-turbo and GPT-4 can effectively detect misinformation, stereotypes, and hate speech without task-specific fine-tuning. High accuracy and F1 scores across datasets validate the robustness of these LLMs as general-purpose language-checking tools.
- **Claim 2-** Unified Prompting Approach: The experiments align with the claim that unified prompting strategies improve performance compared to zero-shot methods. For fact-checking, Few-fp+Zero-cls significantly outperformed Zero-cls, showcasing the efficacy of the unified approach. However, for fairness tasks (Hate Speech & ToxiGen datasets), Zero-cls slightly outperformed the few-shot methods, indicating potential dataset-specific variations.
- **Claim 3-** Grounding Information for Ethical Predictions: The results partially support this claim. While fact-based grounding improved the model's performance for fact-checking tasks, its impact was less pronounced in fairness-checking tasks. This suggests that the inherent grounding within LLMs might already suffice for certain types of ethical evaluations.

- **Claim 4-** Unified Framework for Language Evaluation: The reproduced experiments validate the claim that a single LLM can handle multiple aspects of language evaluation without requiring task-specific models or prompts. The results demonstrate the generalizability and adaptability of the unified framework across fact-checking and fairness-checking tasks.

## 5.1 What was easy

**Dataset Acquisition:** Obtaining the dataset for human-generated text was straightforward since it was readily available in the official GitHub repository.
**Code Availability:** The code for experiments on human-generated text datasets was included in the official repository. The scripts were relatively easy to run, and the explanations in the paper aligned closely with the implementation.
**Baseline Understanding:** The explanation of the zero-shot and few-shot prompting strategies in the paper was well-articulated, making it easier to design additional experiments for these methodologies, even though some parts required new implementation

## 5.2 What was difficult

**Dataset Challenges:** For machine-generated text, the ToxiGen dataset was not included in the official repository, requiring separate download and post-processing, which added time and complexity to the setup. The MGFN dataset was not open-source and could not be obtained, leading to the inability to replicate experiments on the dataset.
**Code Issues:** The code for machine-generated text experiments was not included in the official repository, necessitating additional implementation to reproduce zero-shot and few-shot results. The 'Entailment' methodology was not documented or implemented in the repository. Thus, it couldn't be implemented due to code inavailability, time and resource constraints.
**Technical Challenges:** The official code had outdated API request methods for the OpenAI GPT models, requiring fixes to make proper API calls. Package dependencies and versions, particularly for OpenAI package, were not clearly documented. Resolving compatibility issues and ensuring proper functionality added to the difficulty. The original code also contained issues with logging the execution of the experiments which had to be resolved. The lack of built-in handling for OpenAI API rate limits (e.g., 502 Bad Gateway errors) meant that we had to introduce custom logic, such as adding sleep intervals between prompts, to ensure smooth execution.
**Resource Constraints:**

- The OpenAI GPT-3.5-turbo API's rate limits (10,000 requests per day for Tier-1) caused bottlenecks for larger datasets like SBIC, making it impossible to complete some few-shot experiments.
- Errors such as "Rate limit reached" and "You exceeded your current quota" were frequent, limiting experimentation and increasing execution time.
- The total computational cost incurred for the reproducibility study was $22 of OpenAI credits. This constrained the extent of the experiments conducted.
- Execution time for experiments was significant. Some experiments required over 4 hours to complete zero-shot and few-shot runs on the datasets.

## 5.3 Communication with original authors

While efforts were made to reproduce the experiments using the provided resources, code, and datasets, any unresolved challenges or ambiguities were addressed independently without seeking clarification from the authors.

# References

1. T. Zhang, H. Luo, Y.-S. Chuang, W. Fang, L. Gaitskell, T. Hartvigsen, X. Wu, D. Fox, H. Meng, and J. Glass. **Interpretable Unified Language Checking**. 2023. arXiv: 2304.03728 [cs.CL].
2. T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. **ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection**. 2022. arXiv: 2203.09509 [cs.CL].
3. **GitHub - Vishwa-Sheth/NLP-UniLC: Interpretable unified language safety checking with large language models — github.com**. https://github.com/Vishwa-Sheth/NLP-UniLC. [Accessed 06-12-2024].