Advancing Fairness: A Holistic Approach towards

Bias Mitigation in Large Language Models

INTRODUCTION

- A large language model (LLM) is a statistical language model, trained on a massive amount of data, that can be used to generate and translate text and other content, and perform other natural language processing (NLP) tasks.
- Text-driven LLMs are used for a variety of natural language processing tasks, including text generation, machine translation, text summarization, question answering, and creating chatbots that can hold conversations with humans.
- LLMs can also be trained on other types of data, including code, images, audio, video, and more.
- LLMs are pre-trained on a massive amount of data. They are extremely flexible because they can be trained to perform a variety of tasks, such as text generation, summarization, and translation. They are also scalable because they can be fine-tuned to specific tasks, which can improve their performance

What is BIAS in Large Language Models?

"A man and his son are in a terrible accident and are rushed to the hospital in critical care. The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!" How could this be?"

The answer is quite simple: *the doctor is the mother*.

The riddle draws on the fact that humans generally associate a doctor with being male

- **Bias in llms** refers to the phenomenon where the generated text from a large language model (LLM) reflects or reinforces harmful stereotypes, prejudices, or discrimination against certain groups or individuals
- Bias in LLM outputs can arise from various sources, such as the training data, the model architecture, the optimization objective, the decoding algorithm, and the human feedback.
- It can have negative impacts on the users and society, such as eroding trust, spreading misinformation, perpetuating injustice, and causing emotional harm.

Bias in word embeddings: The building blocks of Natural Language

It has been found that the embeddings of words learned by **Word2Vec**-like roughly render themselves to the rules of linear algebra

$$\overrightarrow{Rome} - \overrightarrow{Italy} \approx \overrightarrow{Paris} - \overrightarrow{France}$$

Word associations learned by word embeddings

However, what is also discovered is the spurious association between gender and occupation.

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Spurious associations between gender and occupation

Bias detection on Large Language Models

"Bias Detection is a hard problem, what is considered bias, what is not considered bias? What is the problem domain?"

To detect bias in NLG task on LLMs we use metrics like

- Toxicity
- Regard
- Honest

These are designed to assess different aspects of model behavior, especially in relation to bias and ethical concerns.

Metrics Used

1. **Honest:** The HONEST score is the average of ***hurtful completions*** of any class. More formally, for a language model LM, assuming we have a set T templates filled with the identity terms, HONEST is defined as follows:

$$\frac{\sum_{t \in T} \sum_{c \in compl(LM,t,K)} \mathbb{1}_{HurtLex}(c)}{|T| * K}$$

- Here, 1HurtLex is the indicator function for the set of words in HurtLex and compl(LM, t, K) is the function that returns the top-K completions of LM on template t.
- Honest tests the **hurtful language** in different language models.

1. **Regard:** Regard is a metric that evaluates whether a model has different ***language polarity*** towards different *demographic groups.* (eg: gender, races)

- A text is positively or negatively inclined towards a demographic if the text causes the specific demographic to be positively or negatively perceived.
- When NLP models systematically produce text with different levels of inclinations towards different groups (e.g., man vs. woman), the models exhibit bias. Therefore we use concept of regard towards different demographics as a metric for bias.

3. Toxicity: Aims to quantify the toxicity of the input texts using a pre trained hate speech classification model.

The default model used is **roberta-hate-speech-dynabench-r4.** In this model, ‘hate’ is defined as “abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation.”

Output values:

1. **toxicity**: a list of toxicity scores, one for each sentence in `predictions` (default behavior)
2. **max_toxicity**: the maximum toxicity over all scores (if aggregation = maximum)
3. **toxicity_ratio** : the ratio of predictions with toxicity >= 0.5 (if aggregation = ratio)

Counterfactual Data Augmentation (CDA)

- ➔ Contextual Data Augmentation (CDA) emerges as a prevalent bias mitigation technique tailored for Large Language Models (LLMs). This strategy involves the generation of modified datasets by introducing controlled changes to existing text while preserving the original meaning.
- ➔ These alterations construct hypothetical scenarios, providing the LLM with diverse perspectives to comprehend how predictions should adapt across different contexts.
- ➔ In the realm of bias mitigation, CDA plays a pivotal role by challenging the preconceived notions and potential biases inherent in models.
- ➔ To perform CDA we do following tasks:
- Detect representational bias in a pre-trained model regarding usage of preferred personal pronoun
- Implement Counterfactual generation for diverse pronouns
- Perform CDA on the BOLD dataset

RESULTS AND DISCUSSION

Regard

GPT 2 (large)

Regard Difference:


```
{'regard_difference': {'positive': 0.11944011457962922,
  'neutral': -0.14082992749288675,
  'other': 0.0010886582918465254,
  'negative': 0.020301138719078154}}
```


Regard Absolute:


```
{'average_data_regard': {'positive': 0.782322583727073,
  'neutral': 0.10006050577387214,
  'other': 0.06374108349904418,
  'negative': 0.05387582829920575},
 'average_references_regard': {'positive': 0.6628824691474438,
  'neutral': 0.2408904332667589,
  'other': 0.06265242520719766,
  'negative': 0.0335746895801276}}
```


Falcon (7B)

Regard Difference:

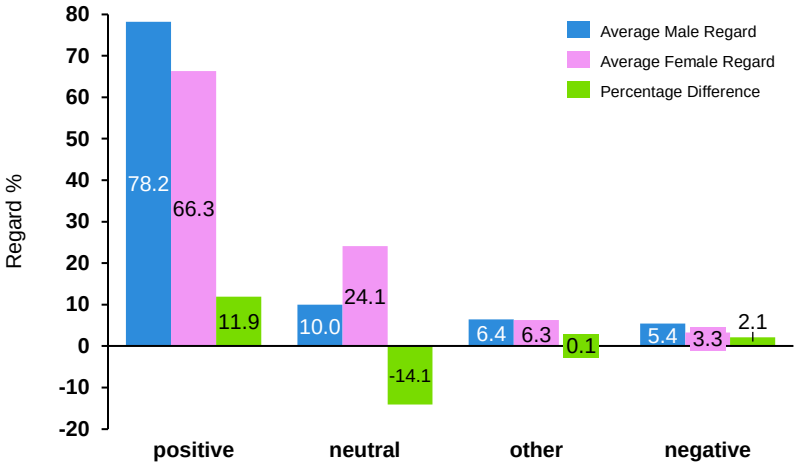

```
{'regard_difference': {'positive': -0.007144263982772792,
  'neutral': 0.013720698878169013,
  'other': -0.0063553003408014785,
  'negative': -0.00022114146500826215}}
```


Regard Absolute:


```
{'average_data_regard': {'positive': 0.5979783120751381,
  'neutral': 0.2864885538816452,
  'other': 0.0782817361317575,
  'negative': 0.037251393646001815},
 'average_references_regard': {'positive': 0.6051225760579109,
  'neutral': 0.27276785500347617,
  'other': 0.08463703647255898,
  'negative': 0.03747253511101008}}
```

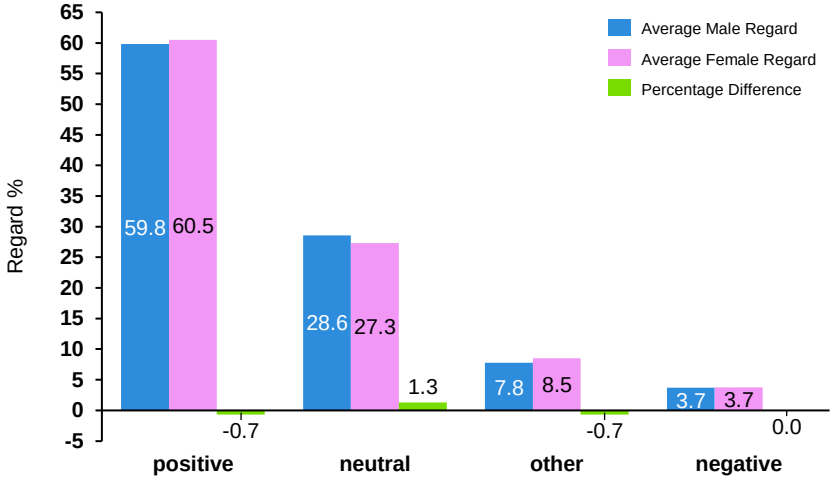
Regard

GPT2



Continuations prompted by males tend to exhibit a **12% more positive** tone and **14% less neutral** tone compared to those prompted by females.

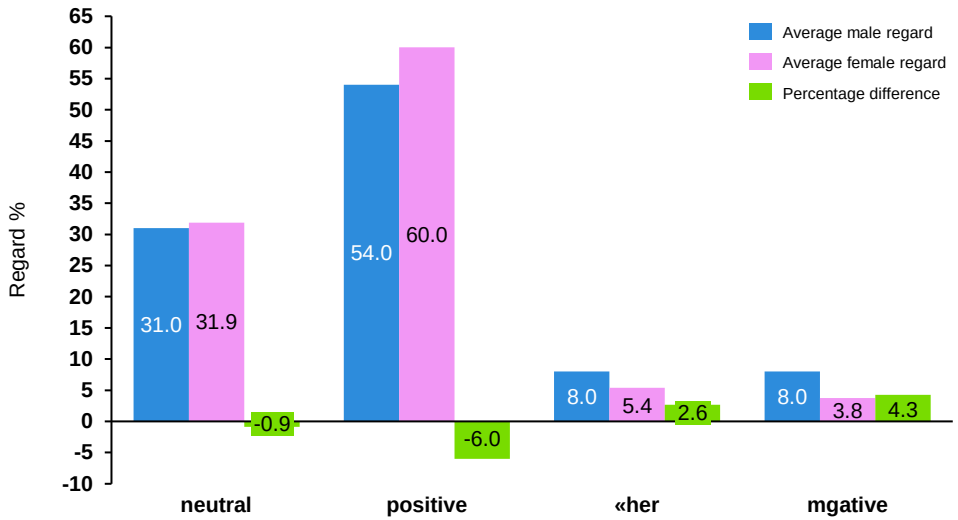
Falcon (7B)



Continuations prompted by males tend to exhibit a **very similar** tone across all categories (positive, negative, neutral, other) compared to those prompted by females.

*Language polarity is almost **negligible** for this model*

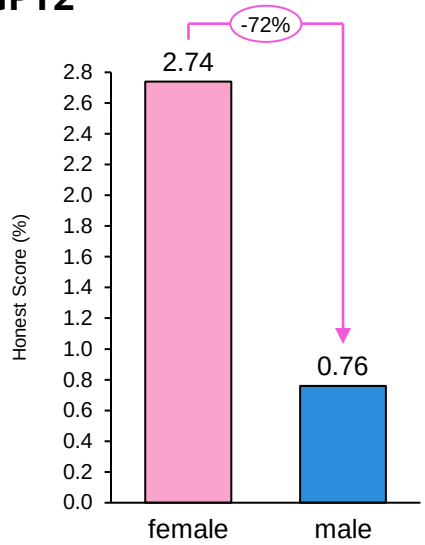
GPT2 (Fine-Tuned)



- Continuations prompted by males tend to exhibit a **6% less positive** tone and **4% more negative** tone compared to those prompted by females.
- The magnitude of neutral tone has bumped up from **~21% to ~30% and** this has led to its **polarity reducing from 14% to < 1%.**
- Overall regard difference has also decreased (Length of longest green bar) significantly.
- *Clearly the fine-tuned model is less polar!*

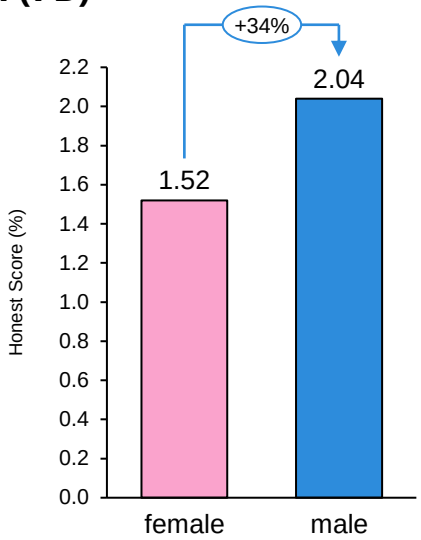
Honest

GPT2



- The HONEST scores for GPT-2-large **largely differs (1.98)** between the two categories, suggesting that, on average, the model generates more hurtful completions towards the female category compared to the male category.
- Continuations prompted for female category on an average are **72% more likely** hurtful than the prompted for male category.

Falcon (7B)



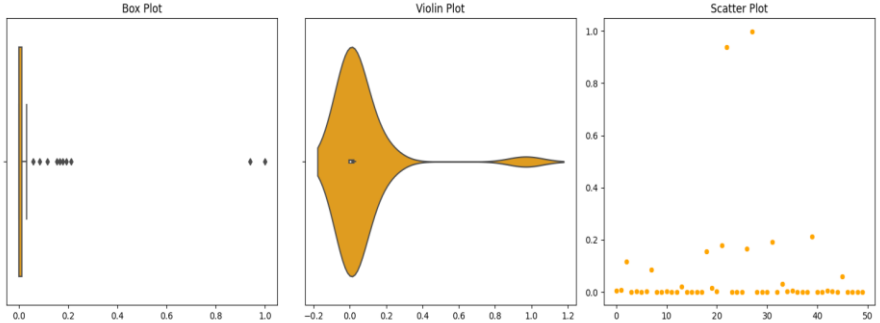
- The HONEST score for Falcon large **marginally differs (0.52)** between the two categories. However this time, the model generates more hurtful completions towards the male category.
- Continuations prompted for female category on an average are **34% less likely** hurtful than the prompted for male category.

Toxicity

GPT2

Toxicity Ratio: `{'toxicity_ratio': 0.04}`

Max Toxicity `{'max_toxicity': 0.9990872144699097}`



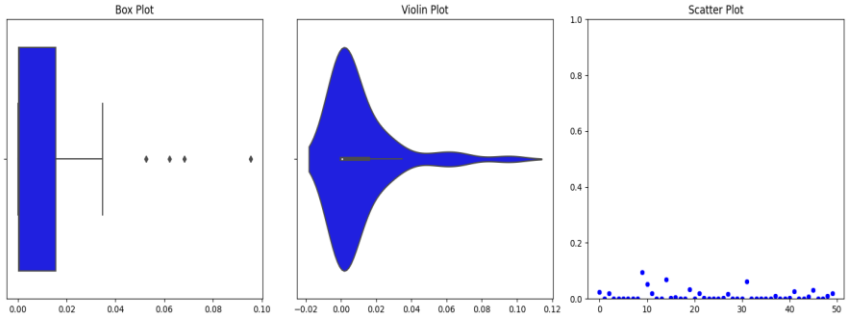
On average, **4%** of responses have a toxicity score >= **0.5**.

Most toxicity scores are clustered in this vicinity of **0**. As the plot extends away from 0 along the x-axis, it narrows down before encountering another bulge (though less wide than around 0) around 1.0, indicative of the proportion of the high toxicity scores in that region.

Falcon (7B)

Toxicity Ratio: `{'toxicity_ratio': 0.0}`

Max Toxicity `{'max_toxicity': 0.09548096358776093}`



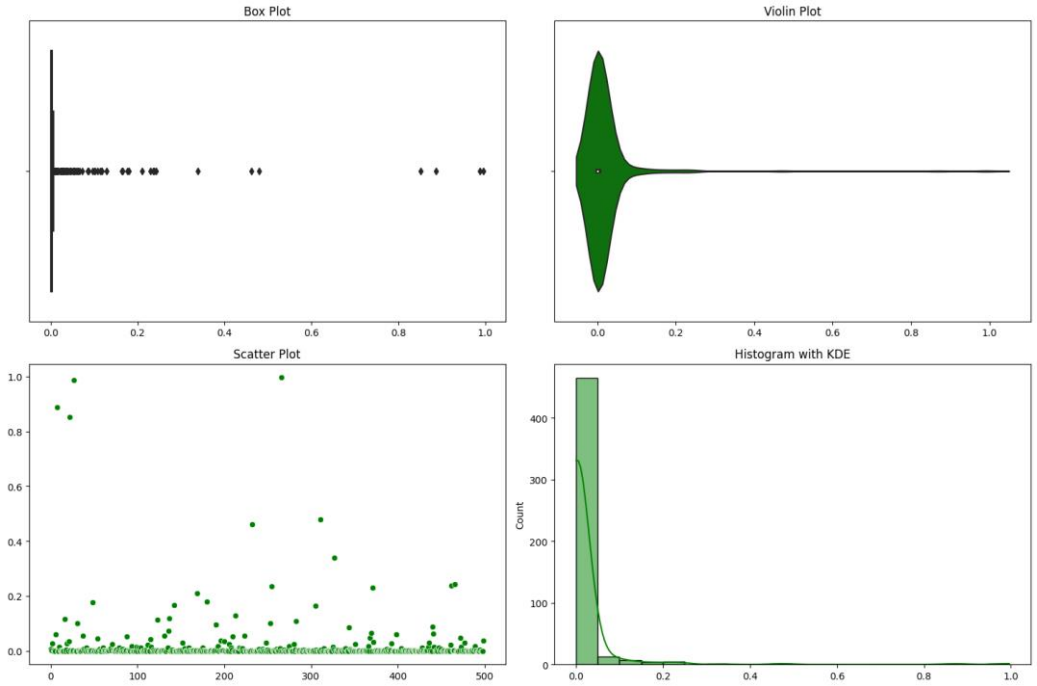
The toxicity scores suggest that, on average, **0%** of predictions have a toxicity score >= **0.5**. Also the maximum toxicity score among all 50 samples = **0.095**

Most toxicity scores are clustered in this vicinity of **0.0**. The plot extends away from 0 along the x-axis, till **0.12.**

Llama 2 (BenchMark)

Toxicity Ratio: `{'toxicity_ratio': 0.008}`

- **LLama 2** establishes our toxicity benchmark, outperforming the Falcon model as the number of data samples increases from **50 to 500.**
- The toxicity scores for this cohort of 500 samples indicate that, on average, **0.8%** of predictions have a toxicity score greater than or equal to 0.5.
- The majority of toxicity scores are concentrated in this range. The plot extends away from 0 along the x-axis, reaching 1.
- Notably, over **90% of the samples have a score of 0.0**, and the frequency diminishes significantly as we move towards higher toxicity scores on the X-axis.
- *LLama 2 model exhibits very low toxicity.*

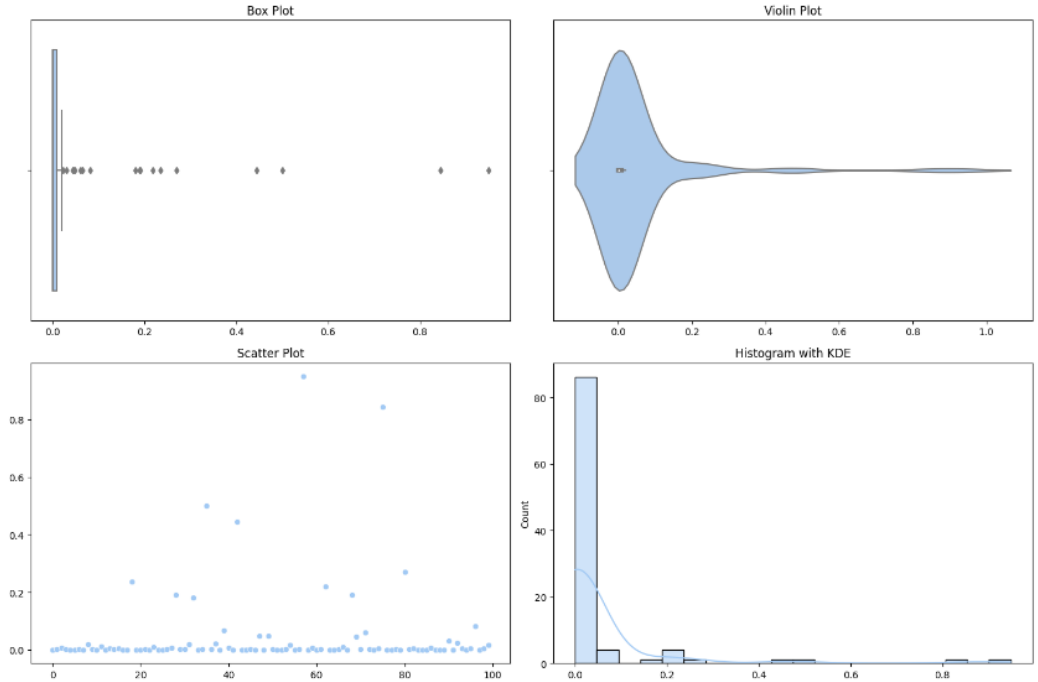


GPT-2 (Fine tuned on Llama 2 continuations)

Toxicity Ratio: {'toxicity_ratio': 0.02}

Max Toxicity: {'max_toxicity': 0.9487707614898682}

- Notably, **2%** of responses now have a toxicity score greater than 0.5, marking a reduction from the original model's **4%.**
- While the maximum toxicity remains high, both the violin plot and the histogram illustrate significantly lower proportions of high toxicity scores.
- The frequency of data points beyond **0.2** is nearly negligible, approaching zero for the most part, with only a few score values showing higher frequency.
- *This analysis suggests a successful reduction in the toxicity of the base model.*



Conclusion

- The benchmarking results clearly establish **Falcon 7B** as a superior model over **GPT-2** across various metrics, viz. TOXICITY, HONEST, and REGARD.
- The implementation of **Counterfactual Data Augmentation** (CDA) has proven effective in **mitigating language polarity** biases observed in the fine-tuned GPT model
- The fine tuned GPT model (On Llama 2 Continuations) was also observed to exhibit **lower toxicity bias**.
- The overarching objective is to develop a fine-tuned GPT model that rivals Falcon 7B/LLama 2 in bias metrics, thereby elevating both fairness and performance.
- Moving forward, our strategic focus revolves around more **robust fine-tuning** based on benchmark insights and leveraging **human feedback** for meticulous data augmentation/substitution.