Fariness in LLM

Sachit
Michigan State University
gaudisac@msu.edu

1. Fariness in LLM

This section aims to delve into the concept of fairness in Large Language Models (LLMs). To investigate fairness in GPT models, we adopt the framework proposed by [2]. Our task involves leveraging generative models for classification on the Adult dataset. We construct natural language queries from the dataset features and utilize next token prediction to classify whether a person will earn more than $50,000. The input prompt is provided in Example.

GPT models struggle with zero-shot learning in generating meaningful next tokens for the task at hand. To address this limitation, we employ few-shot learning by providing the model with curated samples, guiding it to output binary classifications (1 or 0).

To investigate bias in Large Language Models (LLMs), we conduct experiments focusing on the Adult Dataset, addressing simplifications for clarity. Recognizing an inherent imbalance in the dataset ($\times 5.23$), we first balance the occurrences of y=1 and y=0. Given the use of a few-shot data points for guiding predictions, the bias introduced by these few-shot samples significantly influences the query bias. We measure bias using bias parity, denoted as b_{P_c}, calculated as $P(y=1|s=0) - P(y=1|s=1) = 0.1312$. Here, s represents the sensitive attribute (gender in our example), and y indicates income status, where 1 denotes income greater than $50K$, and 0 denotes income less than $50K$. Control over b_{P_c} is achieved by sampling 200 data points according to the specified distribution.

b_{P_c}	ACC	$M_{dpd} \downarrow$	$M_{eod} \downarrow$
0.00	75.5	**0.0049**	**0.0083**
0.13	**85.0**	0.0080	0.0180
0.50	70.5	0.0411	0.0429
1.00	68.5	0.0940	0.1019

Table 1. Few shot performance of GPT models under different bias of the prompts

We can also make the prompt fair using counter balance technique by creating a sample where gender is switched from male to female while maintaining other attributes constant. This adjustment results in a bias parity of 0 for the prompt. We aim to compare this counterbalance approach with an alternative method to achieve $b_{P_c} = 0$, as detailed in Example.

b_{P_c}	ACC	$M_{dpd} \downarrow$	$M_{eod} \downarrow$	Counterbalance
0.00	81.5	**0.0028**	**0.0082**	✓
0.00	75.5	0.0049	0.0083	×

Table 2. Examining Few-Shot Performance of GPT Models: Adding Counterbalanced Samples, i.e explicit changing the gender and keeping everything constant ($b_{P_c} = 0$)

LLMs exhibit fairness challenges and are susceptible to prompt influence. The fairness of predictions aligns with the fairness of the prompt; an unfair prompt tends to result in unfair query predictions.

Given the significant computational resources and time required to train and fine-tune a 7B model, implementing in-processing techniques may pose challenges. However, exploring pre and post-processing techniques offers alternative strategies to address fairness concerns in LLMs.

Some of the ideas to alleviate the problems

- **Adaptive Prompt tuning:** Can prompt tuning be leveraged to discover prompts that yield fair predictions? This concept is akin to the notion of adding noise to images to promote fair predictions. However, the challenge lies in determining whether these prompts generalize across various tasks, ensuring utility across different domains.
- **Training Embedding:** Given the critical role of embeddings, we can explore the concept of MaxEnt-ARL [1]. In this approach, gender information is masked in the prompt, and optimization is performed to maximize entropy for both male and female categories. However, it's is important to note that tuning embeddings may have implications for the model's utility across diverse general tasks.

2. Questions?

Why is the few shot accuracy high?

References

[1] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. 2019. 1

[2] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2023. 1

3. Appendix

Predict whether the annual income of the person is greater than 50k

Text: A person in 1996 has the following attributes:
age 23, workclass Private, education Some-college,
number of years of education 10, occupation Sales,
race Black, sex Female, capital gain 0, capital loss 0,
hours per week 14, native country United-States
label: 0
Text: A person in 1996 has the following attributes:
age 24, workclass Private, education HS-grad,
number of years of education 9, occupation Craft-repair,
race White, sex Male, capital gain 0, capital loss 0,
hours per week 50, native country United-States
label: 0
Text: Text: A person in 1996 has the following attributes:
age 63, workclass Private, education Some-college,
number of years of education 10, occupation Prof-specialty,
race White, sex Male, capital gain 0, capital loss 0,
hours per week 16, native country United-States
label:

Example of input to LLM by sampling two data points as a prompt (few shot =2) and finally one data point from Adult data as a predictor

Predict whether the annual income of the person is greater than 50k

Text: A person in 1996 has the following attributes:
age 23, workclass Private, education Some-college,
number of years of education 10, occupation Sales,
race Black, sex Female, capital gain 0, capital loss 0,
hours per week 14, native country United-States
label: 0
Text: A person in 1996 has the following attributes:
age 23, workclass Private, education Some-college,
number of years of education 10, occupation Sales,
race Black, sex Male, capital gain 0, capital loss 0,
hours per week 14, native country United-States
label: 0
Text: Text: A person in 1996 has the following attributes:
age 63, workclass Private, education Some-college,
number of years of education 10, occupation Prof-specialty,
race White, sex Male, capital gain 0, capital loss 0,
hours per week 16, native country United-States
label:

Example of input to LLM by adding Counterbalance sample, where the datapoint in first and section Text is the same only the gender attribute is changed

b_{P_c}	ACC	$M_{dpd} \downarrow$	$M_{eod} \downarrow$	few shot
0.00	**75.5**	**0.0049**	0.0083	16
0.00	67.0	0.0055	**0.0045**	4

Table 3. Few shot performance of GPT models by varying number of few shot samples.