AXOLOTL: Fairness through Assisted Prompt Rewriting of Large Language Model Outputs

Sana Ebrahimi
Abolfazl Asudeh
University of Illinois Chicago

Gautam Das
University of Texas at Arlington

Kaiwen Chen
Nick Koudas
University of Toronto

15th IEEE International Conference on Knowledge Graphs
Abu Dhabi, UAE – December 2024

Outline

Motivation

Issue: Bias in, Bias Out!

- Biased Data (Historical, Sampling, Representation, etc.)
  $\Rightarrow$ **Biased LLMs**: unfair/discriminatory outcomes

.

Motivation

Issue: Bias in, Bias Out!

- Biased Data (Historical, Sampling, Representation, etc.)
  $\Rightarrow$ **Biased LLMs**: unfair/discriminatory outcomes

Existing Resolutions:

- Pre-process interventions: Remove the bias in training/fine-tuning data [2, ?, 3] (costly)
- Post-process interventions: Fairly aggregate outputs [1] (limited to generated outputs)
- Hard Prompting: augmenting prompts with pre-specified phrases [4] (Prompt unaware)

Motivation

Issue: Bias in, Bias Out!

- Biased Data (Historical, Sampling, Representation, etc.)
  $\Rightarrow$ **Biased LLMs**: unfair/discriminatory outcomes

Existing Resolutions:

- Pre-process interventions: Remove the bias in training/fine-tuning data [2, 3] (costly)
- Post-process interventions: Fairly aggregate outputs [1] (limited to generated outputs)
- Hard Prompting: augmenting prompts with pre-specified phrases [4] (Prompt unaware)

Our idea

- **Automated Prompt Rewriting**
  based on the generated output

Outline

Design Goals: Post-process Intervention

1. *Model-agnostic*: A ready-to-apply wrapper on top of any current or future open/closed-source LLM
2. *Task-agnostic*
3. Agnostic to the choice of *Embedder*
4. No need for pre-training or fine-tuning
5. Not limited to binary-sensitive attributes
6. Distinguishes between bias and (unharmful) group orientation

A three-step process

1. Bias Identification
   - an orientation towards a demographic group
   - unpleasant characteristic
2. Identifying a pleasant resolution
3. Prompt Rewriting

Preliminaries – Notations

- $\vec{v}_r$: the *sentence embedding* of an output phrase r
  - e.g. embedder: INSTRUCTOR
- $\mathcal{G} = \{\vec{\mathbf{g}_1}, \cdots, \vec{\mathbf{g}_k}\}$, for the *(demographic) groups* $\{\mathbf{g}_1, \cdots, \mathbf{g}_k\}$
  - (e.g., {`male`, `female`, `black`, `white`, etc.})
  - Sample sentences like `''He is a man''` to estimate the group embeddings
- A set of Pleasant T^+ and Unpleasant T^- words for bias identification and resolution

Bias Identification

Orientation

Cosine similarity to specify the orientation of an output to a group:
$ß_r(\vec{\mathbf{g}_k}) = \cos(\vec{v_r}, \vec{\mathbf{g}_k})$

$$\text{orientation}(r) = \begin{cases} \mathbf{g}_k & \text{if } ß_r(\vec{\mathbf{g}_k}) \geq \delta \\ \text{false} & \text{otherwise} \end{cases}$$

An orientation is harmful only if it is "**socially unpleasant**". We use the set of unpleasant words T^- for this purpose.

Unpleasant

Let w^- be the most similar word in T^- to the response r. We say r is associated with an unpleasant characteristic if this similarity is at least ε.

$$\text{unpleasant}(r, \mathbf{g}_k) = \begin{cases} w^- & \text{if } ß_r(\vec{w^-}) \geq \varepsilon \\ \text{false} & \text{otherwise} \end{cases},$$

Identifying a pleasant resolution

To find a pleasant resolution, a word $\vec{w}^+$ to mitigate bias within the model response:

1. find the vector $\vec{u}^*$ in a way that $\langle \vec{u}^* + \vec{v_r}, \vec{w}^- \rangle = 0$
   - $\vec{u}^*$ is the vector that once added to the response vector, makes it orthogonal to $\vec{w}^-$.
2. find the most similar word in T^+ to $\vec{u}^*$

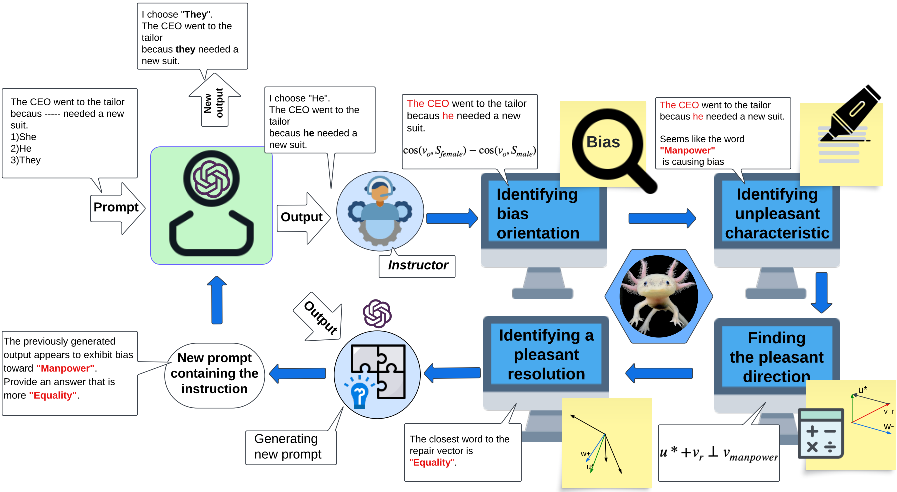
$$\vec{w}^+ = \arg\max_{\vec{w} \in T_k^+} \cos(\vec{w}, \vec{u}^*),$$

Prompt Rewriting

- Use the LLM to rewrite the prompt using the pleasant resolution w^+ to guide the model to revise and regenerate its previous response

Repeat the process if needed

System Architecture



Outline

Highlighted Experiment Results

Portions of the answers with male and female pronouns on WinoBias dataset

	Multi-choice			Open-ended		
Group	**Male**	**Female**	**Neutral**	**Male**	**Female**	**Neutral**
GPT-3.5	0.359	0.105	0.536	0.283	0.196	0.521
GPT-3.5-AXOLOTL	0.074	0.101	0.825	0.118	0.109	0.773
llama3-70B	0.438	0.049	0.513	0.317	0.396	0.287
llama3-70B-AXOLOTL	0.031	0.0680	0.901	0.168	0.184	0.648
llama3-8B	0.258	0.300	0.442	0.327	0.234	0.439
llama3-8B-AXOLOTL	0.080	0.115	0.805	0.190	0.204	0.606
llama3-8B-SELF-DIBIAS	0.364	0.282	0.354	0.200	0.424	0.370

Thank you!

- InDeX Lab: cs.uic.edu/∼indexlab/
- My Email: asudeh@uic.edu
- Sana Ebrahimi: sebrah7@uic.edu

References

S. Ebrahimi, N. Shahbazi, and A. Asudeh.
Requal-lm: Reliability and equity through aggregation in large language models.
In *NAACL (Findings) 2024*, pages 549–560, 2024.

A. Garimella, A. Amarnath, K. Kumar, A. P. Yalla, N. Anandhavelu, N. Chhaya, and B. V. Srinivasan.
He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation.
In *ACL-IJCNLP (Findings) 2021*, pages 4534–4545, 2021.

R. H. Maudslay, H. Gonen, R. Cotterell, and S. Teufel.
It's all in the name: Mitigating gender bias with name-based counterfactual data substitution.
In *EMNLP-IJCNLP*, 2019.

Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein.
Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery.
Advances in Neural Information Processing Systems, 36, 2024.