Requal-LM

Reliability and Equity through Aggregation in Large Language Models

Sana Ebrahimi, Nima Shahbazi, Abolfazl Asudeh

In NAACL 2024 (Findings)

Motivation

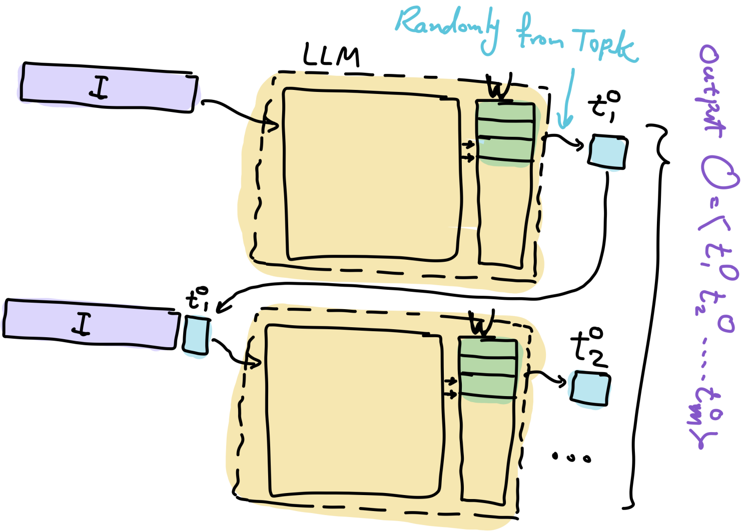
Growing concerns regarding Reliability and Equity in LLM outputs.

- Sequential Randomized nature of LLMs
  - Outputs vary among repeated queries
  - Symmetric tasks where order is not important. E.g., DB queries: shuffling rows should not affect the output
- Inherent biases in data used for training LLMs

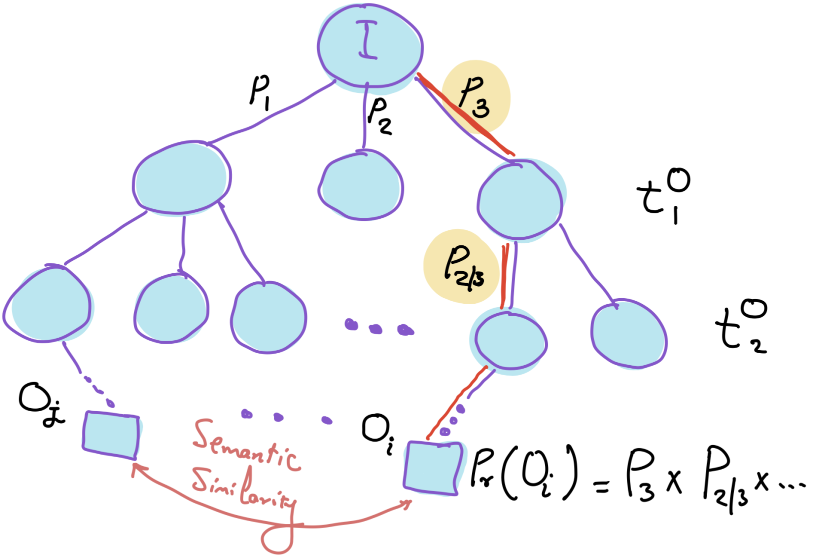
Design Goals

- A ready-to-apply wrapper on top of any current or future open/closed-source LLM
- Task-agnostic
- Agnostic to the LLM of choice and embedder
- No need for pre-training or fine-tuning
- Optimizing both reliability and equity
- Not limited to binary-sensitive attributes
- Distinguishes between harmful and inevitable bias
- Always returns valid results

Randomized Output Generation in LLMs



Output Probability Distribution



Definition

Reliability

Given a prompt I, let

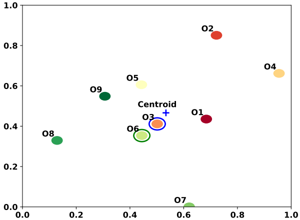
- $\mathcal{O}_I$: universe of possible-to-generate outputs for I
- ξ: the probability distribution of outputs for I ($Pr_\xi(O)$ is the probability that O is generated for I).
- $\vec{\mu}_\xi$: mean of ξ in the embedding space.

Then, the reliability of an output $O_i \in \mathcal{O}_I$ is defined as its similarity to $\vec{\mu}_\xi$.

$$\rho(O) = \mathcal{S}_{im}(\vec{v}_i, \vec{\mu}_\xi)$$

The (unweighted) Monte-carlo method

1. Generate a set of output samples $\{O_1, \cdots, O_m\}$
2. Estimate μ_ξ with the "centroid" of the samples:
$$\vec{v}_c = \frac{1}{m}\sum_{i=1}^{m}\vec{v}_i$$
3. Return the output O_i with the maximum *expected* reliability:
$$\arg\max \quad \big(E\big[\rho(O_i)\big] = \mathcal{S}_{im}(\vec{v}_i, \vec{v}_c)\big)$$



A toy t-SNE of 9 output samples. The green-to-red color code shows the bias values.

Definition

Bias

- Given demographic groups $\mathcal{G} = \{\mathbf{g}_1, \cdots, \mathbf{g}_\ell\}$ and their corresponding vector representation $\{\vec{\mathbf{g}_1}, \cdots, \vec{\mathbf{g}_\ell}\}$. $\Leftarrow$ how?
- Bias of O_i is the maximum similarity disparity of the demographic groups with it. $\beta(O_i) = \max_{\mathbf{g}_j, \mathbf{g}_k \in \mathcal{G}} \left| \mathcal{S}_{im}(\vec{v}_i, \vec{\mathbf{g}_j}) - \mathcal{S}_{im}(\vec{v}_i, \vec{\mathbf{g}_k}) \right|$

Inevitable Bias vs Harmful Bias

- *Inevitable bias*: inherent to the task at hand; not harmful.
$$\beta_n(I) = \min_{O_i \in \mathcal{O}_I} \beta(O_i)$$
- *Harmful bias*: Any bias more than inevitable bias.
$$\beta_h(O) = \beta(O) - \beta_n(I)$$

Objective

Minimize the harmful bias.

The (weighted) Monte-carlo method

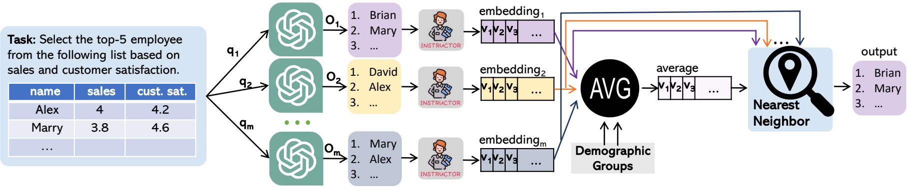
Replace the centroid with the "Equitable centroid ":

- Normalized weight:

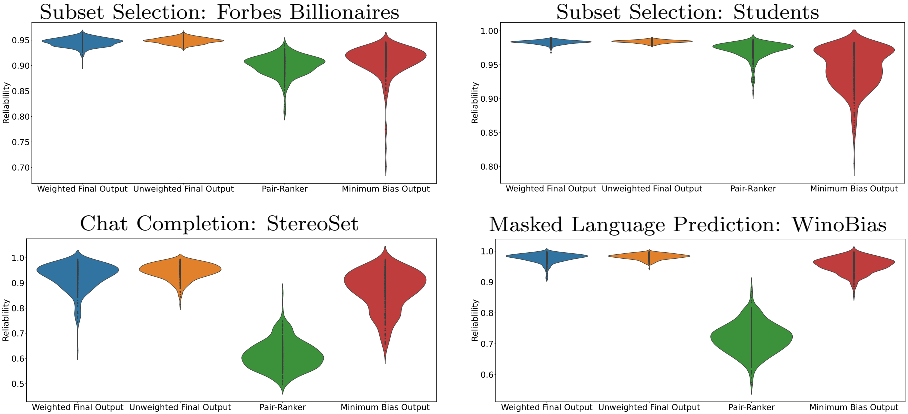
$$w_i = 1 - \frac{\beta(O_i) - \min_{j=1}^{m} \beta(O_j)}{\max_{j=1}^{m} \beta(O_j) - \min_{j=1}^{m} \beta(O_j)}$$

- Equitable Centroid:

$$\vec{v}_c = \frac{1}{m} \sum_{i=1}^{m} w_i \, \vec{v}_i$$



Highlighted Experiments



Highlighted Experiments