2nd April 2025, Lecture Note: GenAI Responsibility: Requal-LM, AXOLOTL

[Harsha Konduru, Yashwanth Reddy]

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Abolfazl Asudeh; www.cs.uic.edu/∼ *asudeh/teaching/archive/cs516spring25/*

Introduction

Large Language Models (LLMs) are at the heart of modern Generative AI (GenAI) systems. They power applications such as chatbots, text summarizers, and content generators. However, these models rely on stochastic generation mechanisms—meaning for the same input prompt, they may generate different outputs each time.

This randomness presents a fundamental challenge in responsible AI: **How do we ensure that the outputs generated are both reliable and fair?**

1 REQUAL-LM: Reliability and Equity through Aggregation

The REQUAL-LM framework, proposes a simple yet powerful solution. By aggregating multiple outputs and using a centroid-based selection strategy, REQUAL-LM improves the reliability and fairness of model outputs in a model-agnostic and task-agnostic way.

1.1 Problem Setting: Randomness in LLM Outputs

LLMs such as GPT and PaLM generate outputs using probabilistic techniques that introduce variability in their responses:

- **Top-k sampling**: Restricts token selection to the k most likely next tokens
- **Top-p (nucleus) sampling**: Selects from the smallest set of tokens whose cumulative probability exceeds threshold p
- **Temperature scaling**: Controls randomness in the output, with higher values producing more diverse responses

For a given input prompt I, each run of the model may produce a different output. Collectively, the set of possible outputs is denoted by $\mathcal{O}_I = \{O_1, ..., O_m\}$. Each O_i is drawn from a probability distribution ξ over the output space $\mathcal{O}_I$.

This inherent randomness raises a critical question: **Which output O_i should we trust and why?** Without a reliable selection strategy, LLM outputs may vary significantly between runs, leading to inconsistent and potentially biased results.

1.2 Design Goals of REQUAL-LM

REQUAL-LM was designed with several key goals in mind to address these challenges:

- **Model-agnostic**: Works on any LLM (even black-box APIs), making it applicable to both open-source and proprietary models
- **Task-agnostic**: Supports a wide range of applications including summarization, question-answering, dialog generation, and more
- **Preprocessing-only**: Requires no model retraining, making it computationally efficient and easy to deploy

- **Valid outputs only**: Filters out hallucinations and irrelevant content, improving output quality
- **Bias-aware aggregation**: Balances reliability with social fairness, addressing both aspects of responsible AI

These design principles make REQUAL-LM highly accessible for real-world deployment without deep technical overhead or specialized resources.

1.3 Reliability through Embedding Similarity

REQUAL-LM defines the reliability of an output using its position in the embedding space relative to other possible outputs. The key insight is that outputs closer to the centroid of the distribution are likely to be more representative and reliable.

For a given output O_i, its reliability is measured as:

- Let $\vec{v}_i$ be the embedding vector of output O_i
- Let $\vec{\mu}_\xi$ be the centroid of embeddings from m samples:

$$\vec{\mu}_\xi = \frac{1}{m}\sum_{i=1}^{m} \vec{v}_i \tag{1}$$

- Then, reliability is defined as:

$$\rho(O_i) = \text{Sim}(\vec{v}_i, \vec{\mu}_\xi) \tag{2}$$

Here, $\text{Sim}(\cdot)$ can be cosine similarity or any other appropriate distance-based metric. The closer an output is to the centroid, the more reliable it is considered to be.

This definition formalizes the intuitive notion that outputs representing the "average" or "mainstream" of possible responses are less likely to contain extreme biases or hallucinations.

1.4 Defining and Measuring Bias

Beyond reliability, REQUAL-LM also addresses bias in LLM outputs. For a set of demographic groups $G = \{g_1, ..., g_\ell\}$ with corresponding vector representations $\{\vec{g}_1, ..., \vec{g}_\ell\}$, the bias of an output O_i is measured as the maximum similarity disparity between demographic groups:

$$\beta(O_i) = \max_{g_j, g_k \in G} |\text{Sim}(\vec{v}_i, \vec{g}_j) - \text{Sim}(\vec{v}_i, \vec{g}_k)| \tag{3}$$

REQUAL-LM further distinguishes between:

- **Inevitable bias**: The minimum level of bias inherent to the task itself

$$\beta_n(I) = \min_{O_i \in \mathcal{O}_I} \beta(O_i) \tag{4}$$

- **Harmful bias**: Any bias beyond the inevitable level

$$\beta_h(O) = \beta(O) - \beta_n(I) \tag{5}$$

This distinction is crucial because some level of bias may be unavoidable for certain prompts, while additional bias represents harmful stereotyping or discrimination that should be mitigated.

1.5 Monte Carlo Aggregation Framework

The core REQUAL-LM method uses a Monte Carlo approach based on repeated sampling to identify the most reliable and least harmfully biased output. The algorithm can be summarized in the following steps:

1.5.1 Unweighted Method (Optimizing for Reliability)

1. Generate m outputs $\{O_1, O_2, ..., O_m\}$ for input I
2. Embed each output to get vectors $\vec{v}_1, ..., \vec{v}_m$
3. Compute the centroid vector $\vec{\mu}_\xi = \frac{1}{m} \sum_{i=1}^{m} \vec{v}_i$
4. Compute the reliability score $\rho(O_i) = \text{Sim}(\vec{v}_i, \vec{\mu}_\xi)$ for each O_i
5. Return the output O_i with the highest $\rho(O_i)$

This approach selects the output closest to the centroid in the embedding space, which is likely to be the most representative of the distribution of possible responses.

1.5.2 Weighted Method (Optimizing for Both Reliability and Equity)

To account for bias, REQUAL-LM introduces a weighted centroid calculation:

1. Compute normalized weights based on bias values:

$$w_i = 1 - \frac{\beta(O_i) - \min_{j=1}^{m} \beta(O_j)}{\max_{j=1}^{m} \beta(O_j) - \min_{j=1}^{m} \beta(O_j)} \tag{6}$$

2. Calculate the equitable centroid using these weights:

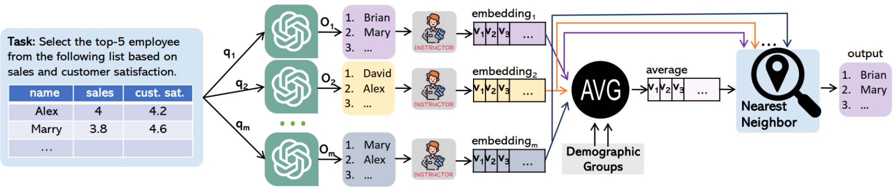
$$\vec{v}_c = \frac{1}{m} \sum_{i=1}^{m} w_i \vec{v}_i \tag{7}$$

3. Return the output closest to this weighted centroid

This weighted approach reduces the influence of biased outputs in determining the final answer, leading to more equitable results without significantly sacrificing reliability.

1.6 System Architecture and Implementation

The REQUAL-LM system architecture consists of several integrated components:



Task: Select the top-5 employee from the following list based on sales and customer satisfaction.

name	sales	cust. sat.
Alex	4	4.2
Marry	3.8	4.6
...		

Figure 1: REQUAL-LM System Architecture.

1. **Input Processing**: Receives the user prompt and prepares it for the LLM
2. **LLM Interface**: Submits the prompt to the LLM multiple times to generate diverse outputs
3. **Embedding Module**: Converts text outputs into vector representations
4. **Centroid Computation**: Calculates the weighted or unweighted centroid of output embeddings
5. **Reliability Scoring**: Measures the similarity between each output and the centroid
6. **Output Selection**: Identifies and returns the optimal output based on reliability and equity metrics

This architecture is designed to be modular and adaptable, working with any combination of LLM and embedding model. The framework can be implemented as a wrapper around existing LLM APIs without requiring access to or modification of the underlying model parameters.

1.7 Experimental Results

REQUAL-LM has been evaluated on tasks like subset selection (e.g., Forbes Billionaires), chat completion, and masked language prediction. The weighted aggregation method consistently outperformed unweighted and baseline approaches by reducing bias while preserving reliability. Notably, it improved female-to-male ratios in selection tasks and increased neutral or anti-stereotypical outputs in co-reference resolution. These results highlight REQUAL-LM's ability to deliver both reliable and fair outputs across varied LLM use cases.

2 AXOLOTL: Fairness through Assisted Prompt Rewriting

Biases in LLMs can manifest in various forms such as gender, race, religion, and profession stereotypes, leading to unfair or discriminatory outcomes in applications ranging from automated hiring systems to conversational AI. Studies have highlighted the critical nature of this problem, demonstrating how biases can skew LLM outputs in ways that reinforce harmful stereotypes and marginalize already disadvantaged groups.

The AXOLOTL framework, named after the Mexican salamander known for its remarkable regenerative abilities, presents a novel post-processing approach to mitigating bias. Just as the axolotl can self-heal and regrow parts of its body, the AXOLOTL framework enables LLMs to identify and correct biases in their own outputs through a process of assisted self-debiasing.

2.1 Problem Setting: Bias in LLM Outputs

The issue of bias in LLMs can be characterized as "Bias in, Bias out." Training data often contains various forms of bias, including historical prejudices, sampling biases, and representation disparities across demographic groups. These biases get encoded into the models and subsequently affect their outputs, leading to unfair or discriminatory outcomes. Researchers have explored multiple strategies to address this problem, each with its own advantages and limitations:

- **Pre-process interventions:** These approaches focus on removing bias from training or fine-tuning data before the model learns from it. While effective, these methods are often computationally expensive and require significant resources.
- **Post-process interventions:** These methods, like REQUAL-LM, aim to fairly aggregate multiple outputs to mitigate bias. While less resource-intensive, they are limited to scenarios where multiple outputs can be generated and aggregated.
- **Hard Prompting:** This technique involves augmenting prompts with pre-specified phrases designed to reduce bias. However, it lacks awareness of the specific biases in each prompt and may not be adaptable to diverse contexts.

AXOLOTL introduces a novel approach: automated prompt rewriting based on the specific biases detected in the generated output. This method offers a more targeted and adaptable solution to bias mitigation.

2.2 Design Goals of AXOLOTL

AXOLOTL is designed as a post-process intervention with the following key characteristics:

- **Model-agnostic:** A ready-to-apply wrapper for any LLM without requiring access to internal parameters
- **Task-agnostic:** The framework is designed to work across various natural language processing tasks.
- **Agnostic to the choice of Embedder:** Flexible with different text embedding models.
- **No need for pre-training or fine-tuning:** AXOLOTL avoids retraining or fine-tuning, making it practical and resource-efficient.

- **Not limited to binary-sensitive attributes:** Handles multiple demographic groups
- **Distinguishes between bias and unharmful group orientation:** AXOLOTL targets only harmful biases linked to unpleasant traits, ignoring benign associations.

These design goals collectively make AXOLOTL a versatile, efficient, and practical solution for addressing bias in a wide range of LLM applications without requiring access to the internal workings of the models themselves.

2.3 Methodology: Three-Step Process

AXOLOTL operates through a three-step process that treats the LLM as a "black box":

2.3.1 Bias Identification

The first step in AXOLOTL's methodology involves identifying two critical components of bias: orientation towards a demographic group and the presence of an unpleasant characteristic. For

orientation detection, AXOLOTL uses vector representations (embeddings) of both the LLM output and the demographic groups of interest. The cosine similarity between these vectors serves as a measure of orientation:

$$\beta_r(\vec{g}_k) = \cos(\vec{v}_r, \vec{g}_k) \tag{8}$$

Where $\vec{v}_r$ is the sentence embedding of an output phrase r, and $\vec{g}_k$ is the vector representation of demographic group g_k (estimated using sample sentences like "He is a man" for male representation).

An orientation is considered significant when this similarity exceeds a predefined threshold δ:

$$\text{orientation}(r) = \begin{cases} g_k & \text{if } \beta_r(\vec{g}_k) \geq \delta \\ \text{false} & \text{otherwise} \end{cases} \tag{9}$$

However, AXOLOTL recognizes that merely having an orientation towards a demographic group doesn't necessarily constitute harmful bias. The orientation becomes problematic when it's associated with socially unpleasant characteristics. To detect this association, AXOLOTL utilizes a set of unpleasant words (T^-) for each demographic group.

The framework identifies the most similar unpleasant word (w^-) to the response and determines if this similarity exceeds another threshold ε:

$$\text{unpleasant}(r, g_k) = \begin{cases} w^- & \text{if } \beta_r(\vec{w}^-) \geq \varepsilon \\ \text{false} & \text{otherwise} \end{cases} \tag{10}$$

This two-part detection mechanism enables AXOLOTL to specifically target harmful biases while allowing benign group associations to remain unaltered.

2.3.2 Identifying a Pleasant Resolution

After detecting a harmful bias, AXOLOTL's second step involves finding a suitable resolution that can guide the debiasing process. This resolution takes the form of a pleasant word (w^+) that can counteract the identified unpleasant characteristic.

The process involves two key operations in the embedding space:

First, AXOLOTL computes a vector $\vec{u}^*$ that, when added to the response vector, would make it orthogonal to the unpleasant characteristic vector:

$$\langle \vec{u}^* + \vec{v}_r, \vec{w}^- \rangle = 0 \tag{11}$$

This orthogonality represents a state where the response no longer aligns with the unpleasant characteristic, effectively neutralizing the harmful bias.

Then, AXOLOTL identifies the word from a set of pleasant words (T^+) whose embedding is most similar to this computed vector $\vec{u}^*$:

$$\vec{w}^+ = \arg\max_{\vec{w} \in T_k^+} \cos(\vec{w}, \vec{u}^*) \tag{12}$$

This pleasant word (w^+) serves as the target direction for rewriting the original response in a way that reduces bias while preserving the essential meaning.

2.3.3 Prompt Rewriting

The final step in AXOLOTL's methodology leverages the LLM's own capabilities to rewrite its output in a less biased way. AXOLOTL formulates a new prompt that includes instructions for debiasing based on the analyses performed in the previous steps.

These instructions typically include information about the detected bias (orientation and unpleasant characteristic) and the suggested pleasant resolution. For example, a prompt might state: "The previously generated output appears to exhibit bias toward 'Manpower'. Provide an answer that is more 'Equality'.

By formulating the instruction in this way, AXOLOTL guides the LLM to self-correct its biases while maintaining the context and purpose of the original response. This approach taps into the model's inherent language understanding and generation capabilities without requiring modifications to its parameters.

If needed, this process can be repeated iteratively, with each new output being analyzed and further refined until a satisfactory level of bias mitigation is achieved.

2.4 System Architecture

The complete AXOLOTL system architecture integrates the three-step process into a cohesive workflow that begins with the initial prompt and ends with a debiased output. This architecture treats the LLM as a "black box," interacting with it only through its standard input and output interfaces.

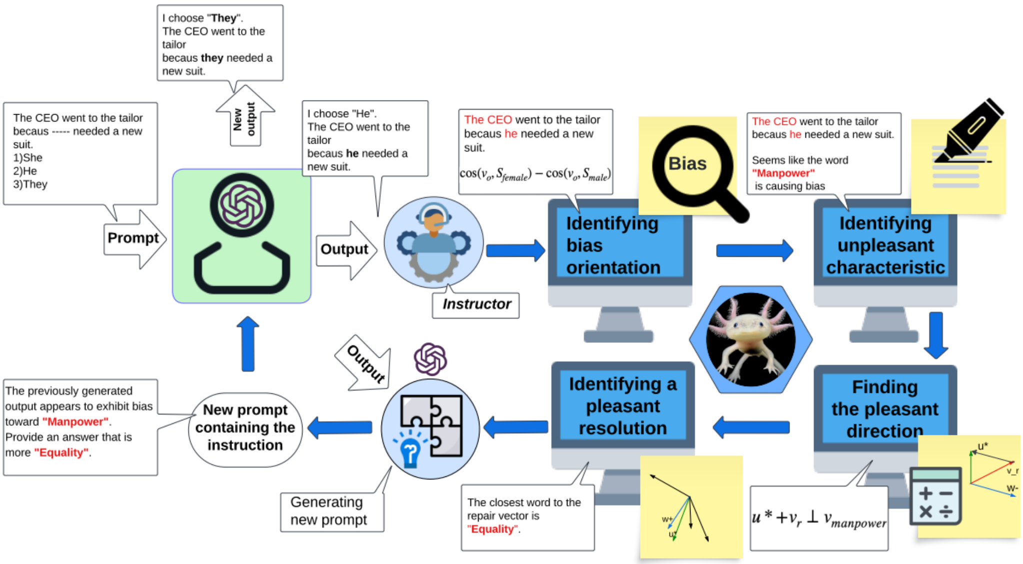

I choose "They". The CEO went to the tailor becaus they needed a new suit.
New output
The CEO went to the tailor becaus ----- needed a new suit. 1)She 2)He 3)They
Prompt
I choose "He". The CEO went to the tailor becaus he needed a new suit.
Output
Instructor
The CEO went to the tailor becaus he needed a new suit. $\cos(v_o, S_{female}) - \cos(v_o, S_{male})$
Identifying bias orientation
Bias
The CEO went to the tailor becaus he needed a new suit. Seems like the word "Manpower" is causing bias
Identifying unpleasant characteristic
Finding the pleasant direction
u*
v_r
w-
$u* + v_r \perp v_{manpower}$
Identifying a pleasant resolution
w+
u*
The closest word to the repair vector is "Equality".
Output
Generating new prompt
New prompt containing the instruction
The previously generated output appears to exhibit bias toward "Manpower". Provide an answer that is more "Equality".

Figure 2: System Architecture

The workflow consists of the following key steps:

1. **Initial Generation:** The LLM produces an initial, possibly biased, response to a user prompt.
2. **Bias Detection:** Embedding analysis identifies orientation toward demographic groups.

3. **Unpleasant Trait Identification:** Detects links to negative characteristics using predefined word sets.
4. **Resolution Calculation:** Computes a repair vector and finds a pleasant word to counteract bias.
5. **Prompt Reformulation:** Generates a new prompt with debiasing instructions.
6. **Debiased Output:** The LLM returns a fairer response based on the revised prompt.

The architecture supports iterative refinement if necessary, with multiple passes through the system until satisfactory bias mitigation is achieved. Throughout this process, the LLM's internal parameters remain unchanged, making AXOLOTL a true post-processing approach applicable to any model.

2.5 Experimental Results

Experiments were conducted on various datasets to evaluate AXOLOTL's effectiveness Using the WinoBias dataset, AXOLOTL significantly increased gender neutralization

Table 1: Proportions of answers with male and female pronouns on WinoBias dataset

Group	Multi-choice			Open-ended		
	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	0.068	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

The results show AXOLOTLimproved gender-neutral responses to over 80%, compared to 50% in original outputs across models.

Conclusion

REQUAL-LM focuses on reliability through embedding-based aggregation, using a Monte Carlo sampling approach to generate multiple outputs and select the most representative one based on centroid proximity. This method effectively reduces hallucinations and biases by filtering out outlier responses that might contain more extreme biases.

AXOLOTL, meanwhile, takes a self-debiasing approach that guides the model to correct its own biases through a three-step process of bias identification, pleasant resolution finding, and assisted rewriting. This method treats the LLM as a black box, making it applicable to any model regardless of architecture or accessibility.

The empirical results demonstrate that both approaches can significantly reduce bias across different tasks and models. REQUAL-LM shows particular strength in enhancing reliability by filtering out anomalous outputs, while AXOLOTL excels at transforming biased outputs into fairer alternatives through guided self-correction.