# Final Exam

Ayesha Ilyas
ailyas6@gatech.edu

## 1 TASK 1

### 1.1 Title

FAIRE: Assessing Racial and Gender Bias in AI-Driven Resume Evaluations

### 1.2 Release date

2 Apr 2025

### 1.3 Link to artifact

https://arxiv.org/abs/2504.01420

### 1.4 Application/scenario/domain of misuse

The FAIRE benchmark ("Fairness Assessment In Resume Evaluation") highlights potential bias in the use of large language models (LLMs) for automated resume scoring in the recruitment and hiring domain. The application scenario involves AI systems that read, evaluate, and score resumes to assist or fully automate hiring decisions. The paper investigates how these models may treat identical resumes differently solely based on demographic characteristics that are inferred by the model like race and gender.

The misuse identified in the FAIRE benchmark involves biased ranking or scoring behavior by AI models, that are potentially disadvantageous to certain demographic groups despite having identical qualifications. This kind of discrimination, even if unintentional, can lead to systemic inequalities in job access and violate fairness and anti-discrimination principles in a critical domain.

### 1.5 Regulated domain/protected class impacted

This artifact use case falls under the employment domain, which is a heavily regulated sector governed by laws like the Civil Rights Act of 1964 (Title VII) and the Equal Employment Opportunity Commission (EEOC) guidelines in the United States. These laws prohibit discrimination based on protected attributes such as race, gender, and national origin.

The protected classes impacted by the AI misuse described in the FAIRE benchmark are:

- Race: The study demonstrates different treatment of resumes based solely on racial cues inferred by the model.
- Gender: Results also show disparities in scoring and ranking when resumes indicate different genders, despite being otherwise identical.

### 1.6 Link to evidence

https://github.com/athenawen/FAIRE-Fairness-Assessment-In-Resume-Evaluation

## 2 TASK 2

**Summary of Bias Identified**

The FAIRE benchmark demonstrates the presence of **algorithmic bias** in large language models used for resume screening. Through controlled experiments that held resume qualifications constant while varying demographic signals (e.g., race and gender), the study uncovered disparities in how resumes were scored or ranked. These disparities constitute violations of **group fairness** principles, including metrics like **Statistical Parity Difference (SPD)** and **Disparate Impact (DI)**, which assess whether different demographic groups receive equitable treatment by an algorithm.

The bias arises from the models' learned associations in training data that may reflect historical discrimination or stereotype reinforcement. As discussed in class, this highlights the critical need for **bias mitigation techniques**, such as **preprocessing algorithms** or **fairness-aware learning**, and aligns with the broader concept of **fairness through awareness**. The artifact illustrates how even advanced AI systems can perpetuate or amplify social inequalities unless fairness is explicitly measured and addressed during model development and deployment.

## 3 TASK 3

1. **Privileged/Unprivileged Groups**: In the FAIRE benchmark, the experiment design explicitly manipulated resumes to reflect variations in race (e.g., White vs. African American names) and gender (e.g., male vs. female pronouns or names). These are legally protected classes under Title VII. Privileged groups include White and Male applicants, while unprivileged groups include African American and Female applicants. The significance lies in observing consistent

outcome differences despite controlled qualifications, suggesting model bias against unprivileged groups.

2. **Sources of Data Bias**: The benchmark's core finding reveals that LLMs often replicate biases present in their training data. Since many LLMs are trained on vast internet corpora (e.g., resumes from job boards or social media), they may reflect societal biases, such as overvaluing resumes associated with certain names or backgrounds. This reinforces harmful stereotypes and violates expectations of neutral scoring.
3. **Sampling Methods Used to Collect Data**: The dataset used in FAIRE was synthetically constructed to control for qualifications while varying protected class signals. This counterfactual sampling approach ensures that any observed differences in model output are not due to qualifications, but solely to protected class attributes. This design strengthens claims of algorithmic discrimination by eliminating confounding variables.
4. **Bias & Fairness Metrics Used to Identify Differences**: The authors evaluated bias using several metrics including **Disparate Impact (DI)** and **Statistical Parity Difference (SPD)**. For example, SPD values closer to 0 and DI values closer to 1 indicate fairness. In several trials, DI scores fell below the 0.8 threshold (the 80% rule), and SPD scores showed significant deviation from 0, indicating unfair treatment of unprivileged groups.
5. **Correlations Found in the Data**: Analysis revealed unintended correlations between race/gender indicators and model output scores. Despite equivalent experience, education, and skills, resumes with African American or female names consistently received lower rankings. This highlights implicit feature importance given to protected-class indicators, which should be orthogonal to merit-based evaluation.
6. **Outcome Measures**: Quantitative results included average rankings and acceptance scores. African American and female applicants were, on average, ranked lower or received lower scores than their privileged counterparts. Standard deviations across groups demonstrated higher variability for unprivileged applicants, which may reflect instability in model decisions. These metrics provide strong statistical evidence of systemic bias.

## 4 TASK 4

One of the key metrics identified in Task 3 is Source(s) of Data Bias, specifically the replication of societal bias within the large language model's training data.

This issue can severely skew predictions when resumes from unprivileged groups (e.g., women or African American applicants) are underrepresented or associated with lower success scores due to biased historical data.

Proposed Strategy: Preprocessing using Counterfactual Data Augmentation (CDA) To mitigate this, we propose a Counterfactual Data Augmentation approach. This strategy involves generating synthetic resume examples by swapping sensitive attributes (e.g., name or gendered language) while keeping qualifications constant. The goal is to neutralize the correlation between protected class attributes and the model's output.

This method aligns with class concepts like Fairness through Awareness and Preprocessing Bias Mitigation.