Bias Detecion in AI Models

Team Name : Codex 2.0

Team lead : Akshita Singh Tyagi

Track : AI Ethics and Bias Mitigation

PROBLEM STATEMENT

Bias Detection in LLMs

1.Bias in Large Language Models (LLMs) can perpetuate discrimination, stereotypes, and unfair treatment.

2.Two primary concerns:

- **Relative Bias:** Different biases exhibited when comparing outputs across multiple LLM
- **Absolute Bias:** Inherent biases within a single LLM's responses.

3.Goal: Detect, analyze, and mitigate biases in LLMs to ensure ethical, fair, and transparent AI applications.

PROPOSED SOLUTIONS

1. Relative Bias Detection

•**Benchmarking with Standardized Prompts:**

- Craft targeted questions covering sensitive topics (e.g., gender, race, socioeconomic factors).
- Compare responses from multiple LLMs (e.g., Llama 3.2, GPT, DeepSeek).

•**Quantitative Metrics Analysis:**

- Use fairness metrics (Demographic Parity, Equal Opportunity).
- Sentiment analysis and toxicity scores for objective bias measurement.

PROPOSED SOLUTIONS

2. Absolute Bias Detection

- **Embedding Analysis:**
  - Examine internal word embeddings for implicit biases.
  - Identify stereotypical associations (e.g anti-national sentiments).

- **Response-based Bias Detection:**
  - Analyze model outputs with specialized bias detection frameworks (GPTBIAS, BiasAlert, FairPy).
  - Detect subtle biases and generate structured reports.

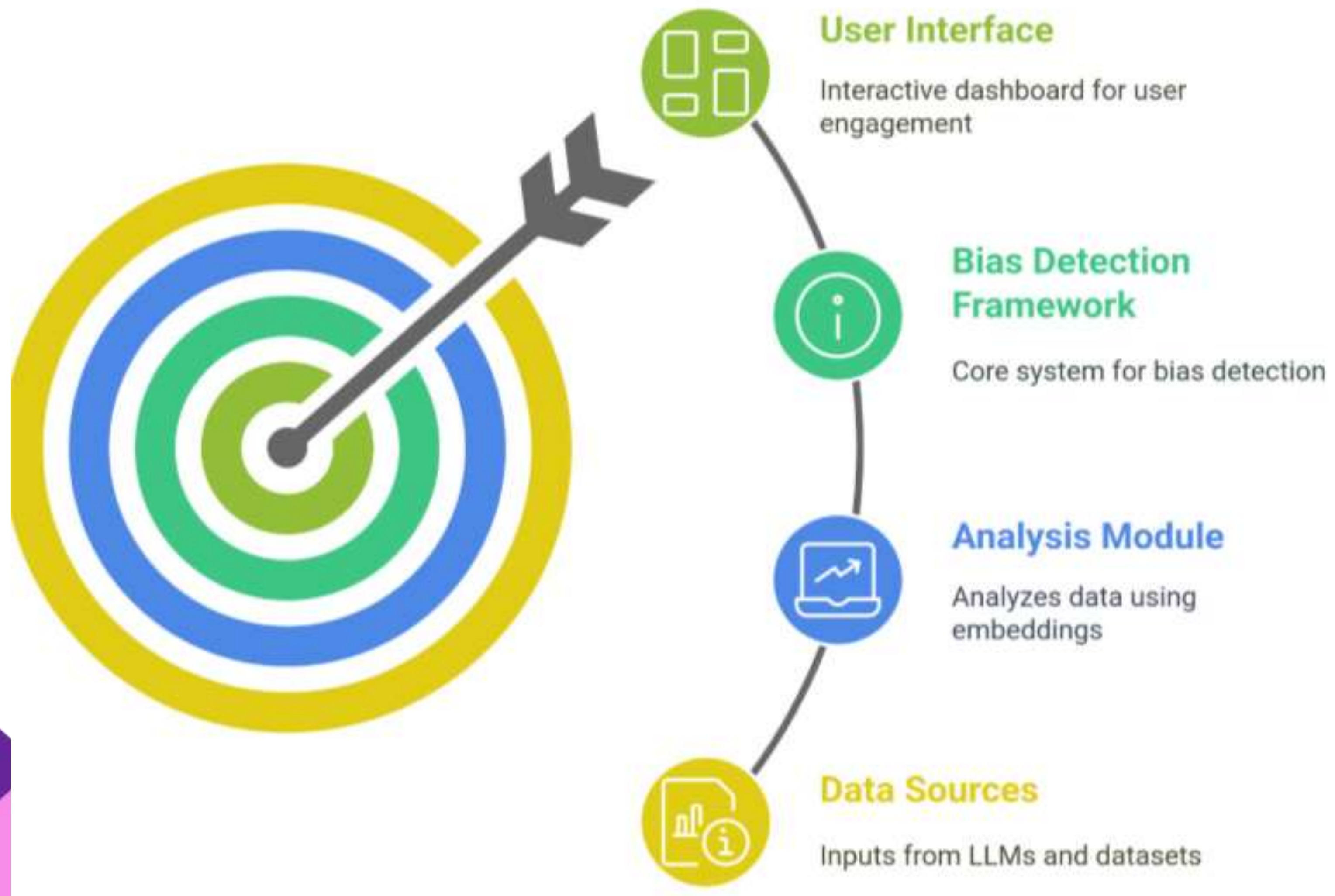
- **Human-in-the-loop Approach:**
  - Combine automated analysis with human evaluation.
  - Validate and interpret nuanced biases.

TECH STACK

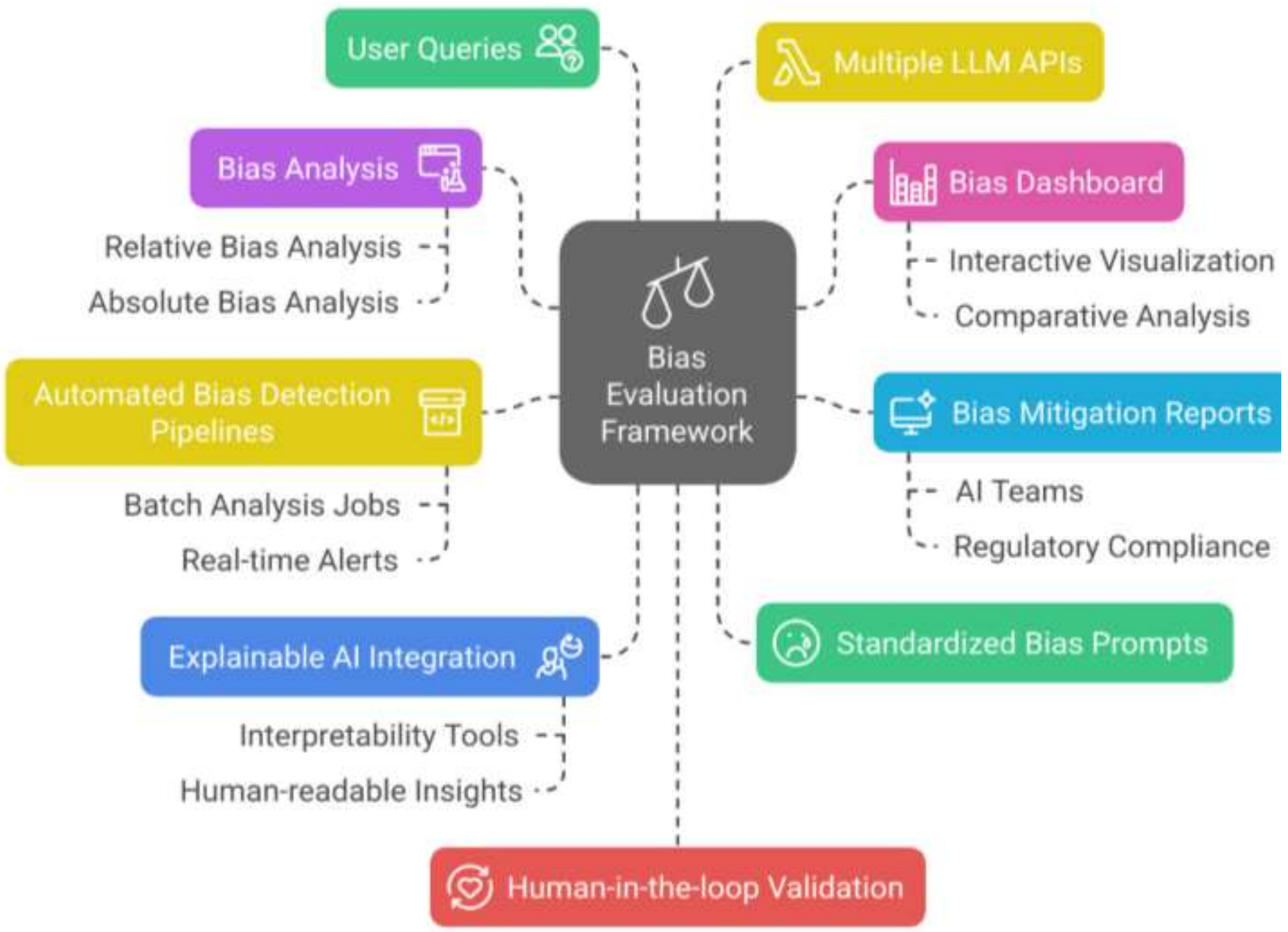
- **Python** (Core implementation).
- **Frameworks:** Fairlearn, AIF360, GPTBIAS, BiasAlert, FairPy.
- **Embedding Analysis:** SpaCy, Gensim.
- **Visualization & reporting:** Matplotlib, Streamlit.

ARCHITECTURE DIAGRAM

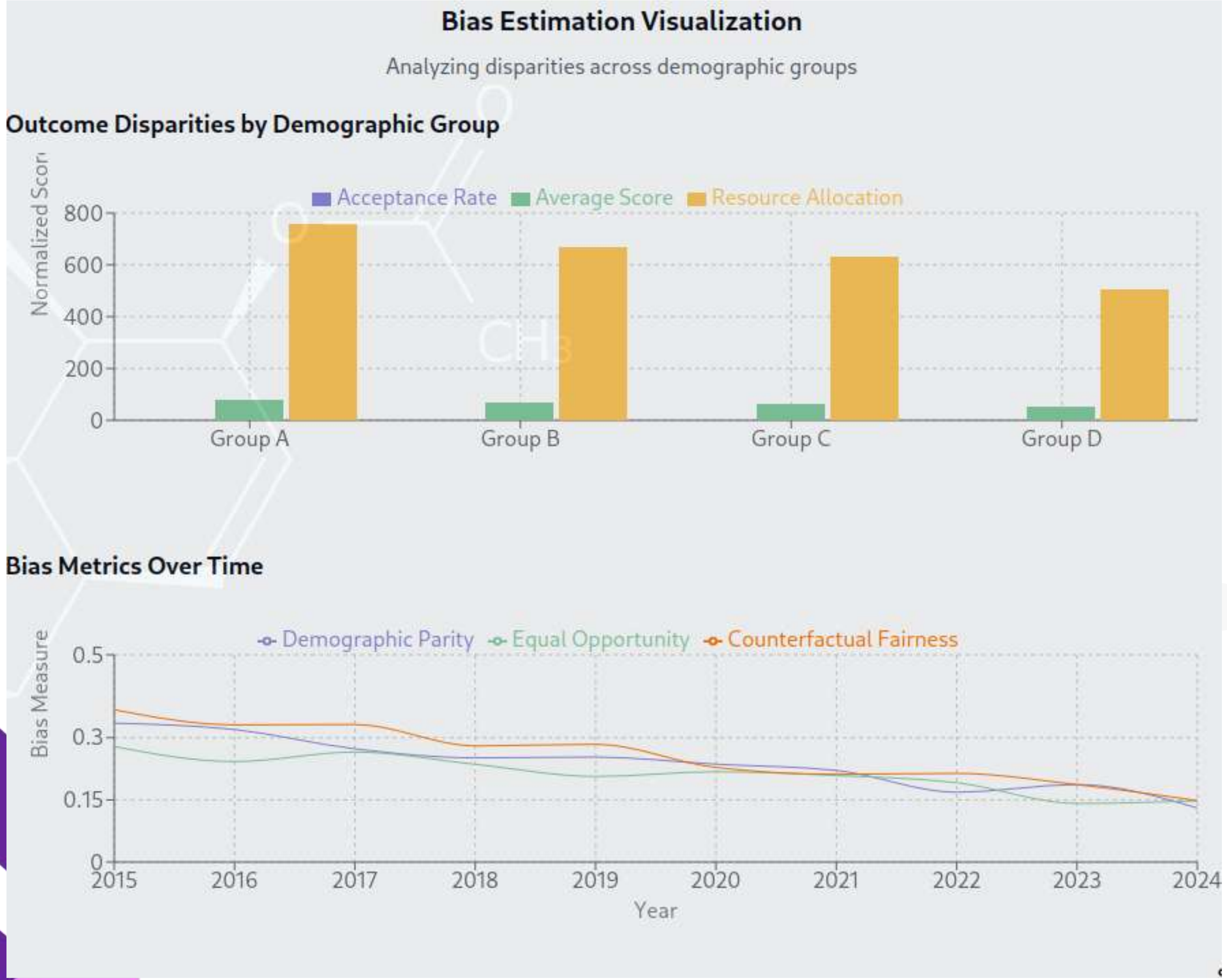
Bias Detection Framework in LLMs



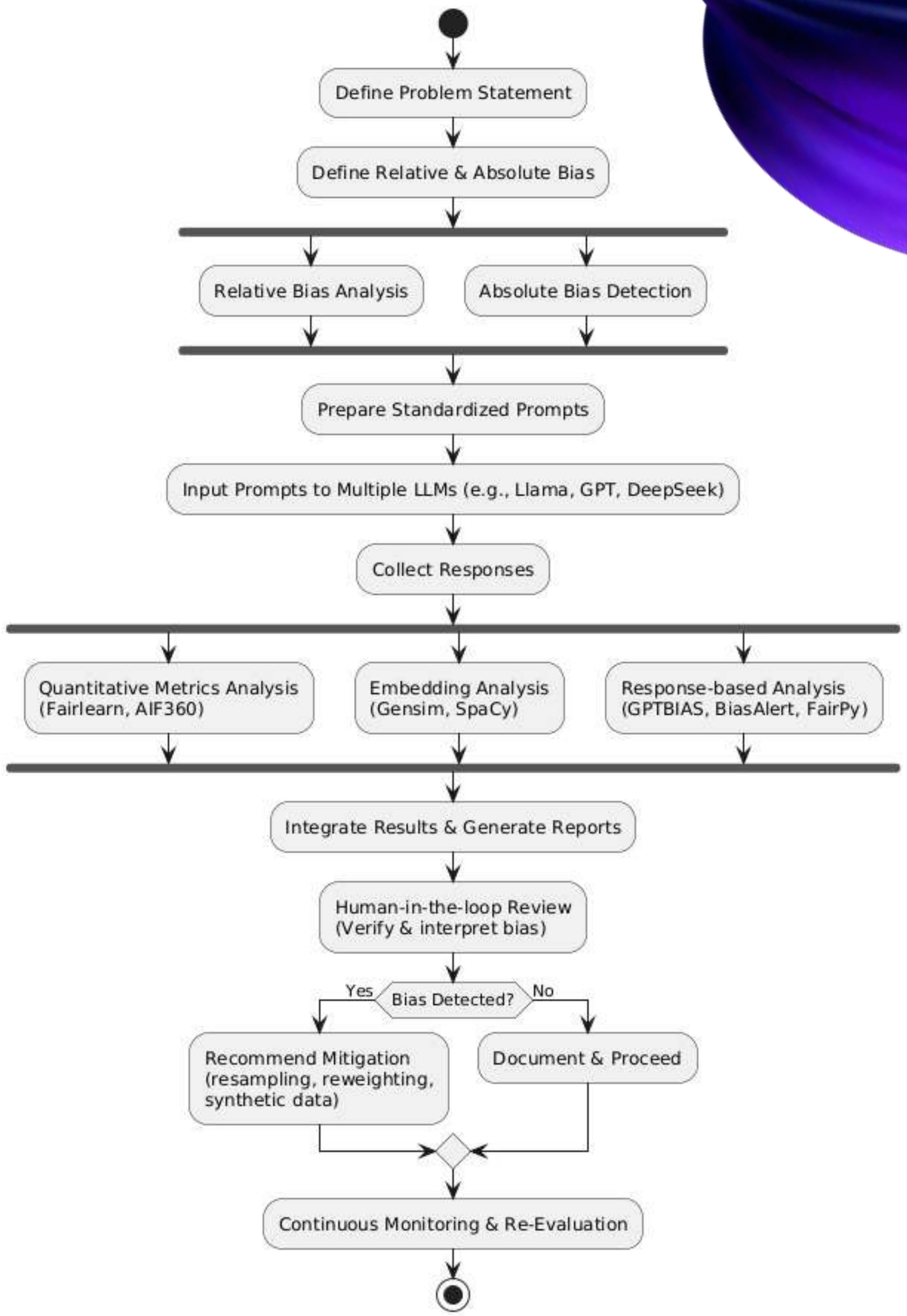
Comprehensive Bias Evaluation Framework for LLMs



GLIMPSES



Bias Estimation Visualization
Analyzing disparities across demographic groups
Outcome Disparities by Demographic Group
Acceptance Rate
Average Score
Resource Allocation
Normalized Scor
800
600
400
200
0
Group A
Group B
Group C
Group D
Bias Metrics Over Time
Demographic Parity
Equal Opportunity
Counterfactual Fairness
Bias Measure
0.5
0.3
0.15
0
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year



Define Problem Statement
Define Relative & Absolute Bias
Relative Bias Analysis
Absolute Bias Detection
Prepare Standardized Prompts
Input Prompts to Multiple LLMs (e.g., Llama, GPT, DeepSeek)
Collect Responses
Quantitative Metrics Analysis (Fairlearn, AIF360)
Embedding Analysis (Gensim, SpaCy)
Response-based Analysis (GPTBIAS, BiasAlert, FairPy)
Integrate Results & Generate Reports
Human-in-the-loop Review (Verify & interpret bias)
Yes
Bias Detected?
No
Recommend Mitigation (resampling, reweighting, synthetic data)
Document & Proceed
Continuous Monitoring & Re-Evaluation

FEASIBILTY AND SCALABILITY

Feasibility:

- **Technical Feasibility:**
  - Proven bias detection frameworks available (Fairlearn, AIF360, GPTBIAS).
  - Existing open-source solutions for embedding and response analysis.
- **Computational Resources:**
  - Moderate computational power needed for response analysis.
  - Higher computational requirements for embedding analysis (GPU resources preferred).

Challenges:

- Requires careful selection and creation of unbiased standardized prompts.
- Interpretation of subtle biases might need expert validation.

FEASIBILTY AND SCALABILITY

- **Impact on AI Deployments:**
  - Ensures ethical and compliant AI applications.
  - Enhances transparency and stakeholder trust.
  - Aligns with emerging AI fairness regulations globally.

- **Scalability Potential:**
  - Methods are generalizable across various LLMs and domains.
  - Automation capability to scale analysis.

- **Target Audience:**
  - AI governance teams, data scientists, regulatory bodies.
  - Organizations deploying AI in regulated industries (healthcare, finance).

- **Benefits:**
  - Reduction in bias (20-50% improvements in fairness metrics).
  - Lower risks related to regulatory non-compliance.
  - Increased stakeholder confidence through transparency and ethical compliance.

TEAM MEMBERS

Team Lead: Akshita Singh Tyagi
RA2311047010075
at3795@srmist.edu.in

Member 1: Ritam Ghosh
RA2311047010055
rg846@srmist.edu.in

Member 2: Ali Khan
RA2211047010066
ak3344@srmist.edu.in

Member 3: Jiya Sehgal
RA2311003011375
js1983@srmist.edu.in