Pre-Launch Bias Audit Framework

Figma-Style One-Page Summary

1. Define Scope & Goals

- Identify model task (e.g., image generation, classification)
- Define intended users & contexts
- Choose fairness dimensions (race, gender, age, disability, intersections)
- Set measurable success criteria

2. Input & Prompt Review

- Build Red Flag Prompt Library (ambiguous/stereotype-prone terms)
- Create Coverage Prompts for multiple groups
- Add Edge Case Prompts for cultural/linguistic stress-tests

3. Model Output Sampling

- Define sample size (≥100 outputs/prompt)
- Randomize seeds for variety
- Collect comparable human baseline data

4. Bias Detection & Quantification

- Human reviewers label outputs for stereotypes/exclusion
- Track representation counts per demographic
- Apply disparity metrics (e.g., demographic parity)
- Create Fairness Scorecard

5. Mitigation Testing

- Experiment with prompt engineering fixes
- Apply data augmentation for balance
- Run adversarial prompts to stress-test model

6. Transparency & Documentation

- Record audit methods + findings in Audit Log
- Update Model Card with limitations + mitigation steps
- Get stakeholder sign-off (Engineering, UX, Ethics, Legal)

7. Post-Launch Monitoring

- Set up user feedback reporting
- Schedule quarterly re-audits or after major updates

Deliverables

- Audit Plan
- Prompt Matrix
- Output Dataset
- Bias Detection Report
- Fairness Scorecard
- Mitigation Recommendations
- Audit Documentation Pack