AI BRIDGE

How-To Guide for Evaluation (Evals) Teams

---

1. Purpose of This Guide

This guide explains **how to run bias evaluations in practice** for the AI-BRIDGE project.

It exists to bridge the gap between:

- Formal evaluation frameworks and metrics
- Early analytical work (fairness definitions, bias labels)
- The **actual, day-to-day evaluation work** done by technical teams

Where this fits in the AI-BRIDGE workflow

This guide is used **after**:

- Data has been collected and annotated
- Ground-truth labels exist
- A bias detection model (or rule-based system) has been implemented

It supports:

- Model comparison
- Fairness validation
- Human-in-the-Loop (HITL) review
- Final evaluation reporting

What evals teams are responsible for

Evals teams are responsible for:

- Measuring model performance (F1 and fairness metrics)
- Comparing multiple modelling approaches
- Identifying where automated outputs fail contextually
- Recording and justifying human overrides

Evals teams are **not** responsible for:

- Redefining bias categories
- Changing annotation guidelines
- Making policy or ethical judgements beyond the defined scope

---

2. How to Use This Guide

When to use it

Use this guide when you are:

- Evaluating a bias detection model
- Comparing multiple models or approaches
- Preparing evaluation results for submission or review

What you need before starting

You should already have:

- A labelled dataset following AI-BRIDGE schemas
- A clear train / validation / test split
- One or more implemented models (rule-based or ML)
- Access to evaluation tooling (e.g. notebooks, MLflow)

Relationship to formal guidelines

- The **formal evaluation documents define what to measure**
- This guide explains **how to measure it and how to interpret results**

---

3. Core Evaluation Principles (Plain Language)

These principles guide all evaluations in AI-BRIDGE:

1. Start simple before going complex

Do **not** jump straight to transformer models. Begin with:

- Rule-based approaches
- Simple baseline models (e.g. logistic regression, KNN, decision trees)

Only move to more complex models once you understand baseline behavior.

2. Evaluation is experimental, not deterministic

You are running **experiments**, not executing a fixed recipe.

This means:

- You do not know the “best” model upfront
- You must compare multiple approaches
- Decisions are based on evidence, not preference

3. F1 score is the primary performance anchor

F1 is used because it balances:

- False positives (over-flagging bias)
- False negatives (missing bias)

A strong F1 score provides the **foundation** for further fairness analysis.

4. Fairness metrics contextualize performance

A good F1 score is **necessary but not sufficient**.

You must also check:

- Whether performance is balanced across genders
- Whether behavior is consistent across languages

5. Human judgement is required

Automated signals are **not final decisions**.

Human reviewers:

- Validate context
- Correct culturally driven misclassifications
- Provide learning signals for model improvement

---

4. Step-by-Step: How to Perform an Evaluation

Step 1: Prepare the data for evaluation

What to look for

- Correct eval split (test set only)
- Complete required fields (language, bias label, severity, etc.)
- PII removed and safety flags respected

What to do

- Confirm schema consistency
- Exclude rejected or “needs_review” samples
- Lock the test set before evaluation

What not to do

- Do not tune models on the test set
- Do not change labels during evaluation

Common mistakes

- Mixing validation and test data
- Evaluating on partially annotated samples

Step 2: Run baseline models first

What to look for

- Performance of simple approaches
- Clear strengths and weaknesses

What to do

- Train multiple baseline models (e.g. logistic regression, decision trees)
- Record F1 score for each
- Compare results side-by-side

What not to do

- Do not skip baselines
- Do not assume transformers will perform better by default

Common mistakes

- Over-engineering too early
- Lack of comparative evidence

Step 3: Track experiments consistently

What to look for

- Reproducibility
- Metric history over time

What to do

- Use experiment tracking (e.g. MLflow)
- Log:
    - Model type
    - Parameters
    - F1 score
    - Fairness metrics

What not to do

- Do not rely on memory or screenshots
- Do not overwrite previous runs

Common mistakes

- Losing experiment history
- Inconsistent metric naming

Step 4: Evaluate fairness metrics

What to look for

- Disparities between groups or languages

What to do

- Calculate:
  - Demographic Parity (DP)
  - Equal Opportunity (EO)
  - Average Odds Difference (AOD)
  - Multilingual Bias Evaluation (MBE)
- Compare results across groups

What not to do

- Do not treat fairness metrics as pass/fail in isolation
- Do not ignore context behind disparities

Common mistakes

- Reporting metrics without interpretation
- Ignoring multilingual inconsistencies

Step 5: Apply Human-in-the-Loop review

What to look for

- False positives caused by cultural context
- False negatives where bias is subtle or implicit

What to do

- Conduct expert review on flagged samples
- Measure:
  - Human–Model Agreement Rate (HMAR)
  - Annotation consistency (Kappa / Alpha)
- Record rationale for overrides

What not to do

- Do not override silently
- Do not treat human review as optional

Common mistakes

- Inconsistent human decisions
- Poor documentation of judgement calls

Step 6: Decide whether issues are material

What to look for

- Systematic patterns, not one-off errors
- Bias severity and recurrence

What to do

- Assess whether issues:
  - Affect specific genders or languages repeatedly
  - Impact model credibility
- Escalate material risks

What not to do

- Do not escalate isolated edge cases
- Do not ignore repeated small issues

Common mistakes

- Overreacting to noise
- Underreacting to patterns

Step 7: Document results

What to look for

- Clarity and traceability

What to do

- Produce a Bias Audit Report including:
  - Model comparisons
  - Key metrics
  - Human overrides and rationale
  - Known limitations

What not to do

- Do not submit raw metrics without explanation

Common mistakes

- Vague summaries
- Missing justification for decisions

---

5. Handling Grey Areas and Edge Cases

When rules are not clear

- Default to **expert ML guidance**
- Use consistency over perfection

When automated signals conflict with human judgement

- Human judgement takes precedence
- The conflict must be documented

When to escalate

Escalate when:

- Bias affects a protected group systematically
- Metrics degrade across multiple languages
- There is uncertainty the team cannot resolve

How to stay consistent

- Use shared evaluation channels
- Align on examples and interpretations
- Document decisions transparently

---

6. Examples

Only include examples already supported by:

- Evaluation datasets
- Expert walkthrough notebooks
- Prior team evaluations

If examples are not explicitly validated, **do not invent new ones**.

---

7. Documentation & Traceability

What must be recorded

- Model versions
- Evaluation metrics
- Human review outcomes
- Rationale for overrides

Why this matters

- Supports auditability (Gates Foundation requirements)
- Enables learning across iterations
- Protects teams from ambiguity later

What “good enough” looks like

- Clear, concise reasoning
- Repeatable steps
- No unexplained decisions

---

8. What This Guide Does Not Cover

This guide does not:

- Define new fairness metrics
- Redesign bias taxonomies
- Replace gender expert review
- Act as an ethics or policy document

---

9. Glossary of Terms (Operational)

F1 Score

Balanced measure of precision and recall used as the primary performance metric.

Demographic Parity (DP)

Checks whether predictions are evenly distributed across groups.

Equal Opportunity (EO)

Compares true positive rates across groups.

Average Odds Difference (AOD)

Measures combined disparity in true and false positive rates.

MBE (Multilingual Bias Evaluation)

Assesses fairness consistency across languages.

Human–Model Agreement Rate (HMAR)

Percentage agreement between model predictions and human judgement.

Human-in-the-Loop (HITL)

Structured process where human experts validate and correct model outputs.