Ten-Week Research Roadmap

Detecting Implicit Bias in LLM Outputs

Prepared for David

February 11, 2026

Abstract

This document specifies a concrete, ten-week plan (five two-week milestones) to detect *implicit* bias in the outputs of large language models (LLMs). It fixes the primary evaluation dataset, enumerates exact ablation studies, defines bias-detection metrics (and the underlying *matrix*-based counting structure), and provides clear deliverables, risks, and reproducibility steps. The plan is suitable for rapid execution by a single researcher and aims toward a workshop-quality manuscript by Week 10.

Contents

1 Scope & Research Questions

Goal. Quantify and analyze implicit social biases exhibited by instruction-tuned LLMs under realistic prompting conditions, and evaluate mitigations via prompt and decoding interventions.

Core questions.

1. When presented with ambiguous vs. disambiguated contexts, how often do LLMs choose stereotype-consistent outputs?
2. Which controllable factors (prompting style, few-shot context, decoding hyperparameters, persona/system prompts) most affect bias scores?
3. How robust are bias estimates across demographic axes and templates?

2 Primary Dataset (Fixed)

Primary dataset (use this specifically).

BBQ (Bias Benchmark for Question Answering)

BBQ v1.0 (2022): A balanced QA benchmark targeting social bias across many demographic axes with paired *ambiguous* and *disambiguated* contexts. Use the multiple-choice format to measure stereotype reliance vs. evidence-based answering; report primary results on BBQ.

Optional validation sets (stress-tests, not mandatory): CrowS-Pairs (minimal pairs), StereoSet (LM preference), HolisticBias (broad descriptors), BOLD (open-ended prompts). These are supplementary; the *primary* figures will be from BBQ.

3 Milestone 1 (Weeks 1–2): Literature Review

Objectives

- Read **12-15 core papers** in-depth (close reading with notes).
- Maintain an **Excel sheet** with a **5–7 sentence summary of at least 15 *other* papers**.

Core Reading List (You can also choose other papers if you find any relevant ones.)

1. BBQ: Bias Benchmark for QA (2022).
2. CrowS-Pairs (2020).
3. StereoSet (2021).

4. WinoBias (2018).
5. WinoGender (2018).
6. RealToxicityPrompts (2020).
7. BOLD (2021).
8. Bias in Bios (2019).
9. The Woman Worked as a Babysitter? (2019).
10. Language (Technology) is Power (2020).
11. Stochastic Parrots (2021).
12. HolisticBias (2022).

Excel Tracking (for 15+ additional papers)

Create a spreadsheet with the following columns:

Paper Title	**Year**	**5–7 Sentence Summary**	**Relevance (1–5)**

Deliverables

- A 2–3 page review memo (key findings, open problems).
- Completed Excel with 15+ additional paper summaries.

Risks & Mitigations

- *Risk:* Too broad scope. *Mitigation:* Keep BBQ as the fixed primary.

4 Milestone 2 (Weeks 3–4): Dataset Collection & Assets

Objectives

- Acquire **BBQ v1.0**. Create `train/dev/test` splits with preserved group balance.
- Prepare prompt templates mirroring BBQ question styles (*ambiguous* vs *disambiguated*).
- (Optional) Ingest CrowS-Pairs / StereoSet for validation-only.

Concrete Steps

1. **Data card.** Document axes, licensing, and intended use.
2. **Normalization.** Ensure consistent tokenization; lowercase-only and punctuation-normalized variants for sensitivity checks.
3. **Counterfactual pairs.** For each demographic term, create *swap* templates (e.g., Group A $\leftrightarrow$ Group B) to support controlled comparisons.
4. **Prompt library.** Store all `.txt` templates in `prompts/` with fields for `{context}`, `{question}`, `{options}`.

Deliverables

- `data/` with BBQ splits; `prompts/` with templates; `README.md` data card.

5 Milestone 3 (Weeks 5–6): Models & Ablation Studies

Baseline Models

- **Primary:** An instruction-tuned open LLM (e.g., 7–13B class).
- **Secondary:** A larger instruction-tuned open LLM (e.g., 30–70B class) for scaling comparisons.

Exact Ablations (one factor at a time)

A1. **Prompting style:** zero-shot vs. few-shot ($k \in \{1, 3, 5\}$).

A2. **Rationale prompting:** direct answer vs. chain-of-thought style (when allowed) vs. *brief justification.*

A3. **Demographic cue salience:** explicit (e.g., group labels) vs. implicit (e.g., names) vs. removed.

A4. **Instruction header:** neutral system prompt vs. de-bias reminder ("Base your answer only on evidence; avoid stereotypes").

A5. **Decoding:** greedy vs. temperature $\in \{0.0, 0.3, 0.7\}$; top-$p \in \{0.9, 0.95\}$.

A6. **Example selection:** random vs. *anti-stereotype* exemplars in few-shot context.

A7. **Position of cues:** demographic terms early vs. late in the prompt.

A8. **Length control:** max-tokens $\in \{16, 64\}$ to test verbosity effects.

A9. **Model size:** small vs. large instruction-tuned models.

A10. **Safety layer:** on vs. off (if applicable for the chosen stack) to observe filtering effects.

Implementation Notes

- For multiple-choice BBQ, compute *option log-likelihoods* and pick $\arg\max$; record margins for analysis.
- Cache tokenization and logprobs; set random seeds; run each condition with 3 independent seeds.

Deliverables

- `ablations.csv` with columns: `model, k, rationale, cue_type, instr, temp, top_p, pos, max_tokens, safety, seed, metric...`

6 Milestone 4 (Weeks 7–8): Evaluation & Bias Detection

Bias Detection via a *Matrix* of Counts

For each demographic group g, construct a 1×3 row of counts:

$$\text{Counts}_g = \left[n_{g,\text{stereo}},\; n_{g,\text{anti}},\; n_{g,\text{neutral}}\right].$$

Stacking across two groups (A,B) forms a **Bias Confusion Matrix**:

$$\mathbf{B} = \begin{bmatrix} n_{A,\text{stereo}} & n_{A,\text{anti}} & n_{A,\text{neutral}} \\ n_{B,\text{stereo}} & n_{B,\text{anti}} & n_{B,\text{neutral}} \end{bmatrix}.$$

This matrix underlies all metrics below.

Primary Metrics (BBQ)

M1. **Accuracy (ACC):** fraction of correct answers (higher is better).

M2. **Stereotype Reliance Rate (SRR):** proportion of *incorrect* answers that are stereotype-consistent:

$$\text{SRR} = \frac{\sum_g n_{g,\text{stereo-wrong}}}{\sum_g n_{g,\text{wrong}}}.$$

M3. **Bias Score (BBQ-style):** mean preference for stereotype-consistent choice on *ambiguous* items minus *disambiguated* items (normalized to $[-1, 1]$).

Group-Fairness Metrics

F1. **Demographic Parity Gap (DPG):** $|\text{ACC}_A - \text{ACC}_B|$.

F2. **Equalized Odds Gap:** average of TPR and FPR differences between groups.

F3. **Calibration Gap (Brier):** $|\text{Brier}_A - \text{Brier}_B|$ when using option probabilities.

Log-Probability Preference (for validation sets)

For minimal pairs $\langle s_{\text{stereo}}, s_{\text{anti}} \rangle$, define the **stereotype logprob gap**

$$\text{SLG} = \frac{1}{N} \sum_{i=1}^{N} \Big(\log p(s_{\text{stereo}}^{(i)}) - \log p(s_{\text{anti}}^{(i)}) \Big).$$

Positive SLG indicates stereotypical preference.

Uncertainty & Significance

- Nonparametric **bootstrap** (5,000 replicates) for CIs on ACC, SRR, DPG.
- **McNemar's test** on paired choices across ablations.
- Report **effect sizes** (Cohen's h for proportions).

Visualizations

- Bias confusion matrix heatmaps by group.
- SRR vs. ACC trade-off plots across ablations.
- Reliability diagrams per group (when probabilities available).

Deliverables

- `results/` with per-item logs; `figures/` (PDF/PNG) for all plots; `tables/` with summary metrics and CIs.

7 Milestone 5 (Weeks 9–10): Manuscript Writing

We will guide you for this when you come to this stage.

A Appendix A: Excel Columns (Copy/Paste)

```
Columns:
- Paper Title
- Year
- Venue
- Problem / Task
- Dataset(s) Used
- Method Summary (5-7 sentences)
- Key Metrics
- Key Results
- Limitations / Critique
- Relevance (1-5)
```

B Appendix B: Metrics (Formal)

Accuracy (ACC). $\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}$.

Stereotype Reliance Rate (SRR). Let $\mathcal{E}$ be the set of errors. For each error, classify as stereotype-consistent or anti-stereotype using BBQ labels:

$$\mathrm{SRR} = \frac{1}{|\mathcal{E}|}\sum_{i\in\mathcal{E}} \mathbf{1}\{\text{stereotype-consistent}\}.$$

Demographic Parity Gap (DPG). $|\mathrm{ACC}_A - \mathrm{ACC}_B|$.

Equalized Odds Gap. $\frac{1}{2}(|\mathrm{TPR}_A - \mathrm{TPR}_B| + |\mathrm{FPR}_A - \mathrm{FPR}_B|)$.

Brier Score (grouped). $\mathrm{Brier}_g = \frac{1}{N_g}\sum(\hat{p}_i - \mathbf{1}\{\hat{y}_i = y_i\})^2$.

SLG (minimal pairs). $\mathrm{SLG} = \frac{1}{N}\sum_{i=1}^{N}\Big(\log p(s_{\text{stereo}}^{(i)}) - \log p(s_{\text{anti}}^{(i)})\Big)$.

C Appendix C: File/Folder Layout

```
project-root/
  data/
    bbq/ (splits, README)
  prompts/
    bbq_ambiguous.txt
    bbq_disambiguated.txt
  eval/
    run_eval.py
    metrics.py
    bias_matrix.py
  results/
    logs/
    tables/
    figures/
  ablations.csv
  README.md
```