Pareto-Efficient Debiasing via Causal Localization and Local Circuit Edits (CMA × SFC-lite)

Haoyang Yin, Qiji Zheng, Qiao Zhao

1. Motivation & Problem Statement

A common way to reduce stereotype bias in language models is to ablate whole attention heads or even layers. This often helps fairness but causes collateral damage (perplexity ↑, downstream accuracy ↓). By contrast, causal localization (e.g., CMA/mediation in Vig et al., 2020) pinpoints which units carry bias but does not specify how to edit them with minimal performance degradation.

We propose a closed-loop procedure:

- Localize a small set of causally responsible units (Top-K heads/channels) via CMA.
- Validate local paths within this set and perform fine-grained edits (cuts or soft gates) — SFC-lite (cf. Marks et al., 2025, "Sparse Feature Circuits") at the feature/connection level without global circuit mining.
- Evaluate at matched debiasing, plotting the Bias–Perplexity Pareto frontier; additionally, for each mediator we quantify the dataset-level performance impact of single-site replacement (NIE vs. ΔPerplexity).

Goal: achieve the same debiasing (matched TE/NIE reduction) with smaller performance cost and smaller structural change than head ablation.

2. Research Questions

- RQ1 — Causal localization & minimal cuts. Can a small mediator set explain the bias, and can we extract a minimal cut that substantially reduces TE/NIE?
- RQ2 — Pareto at matched debiasing. At matched TE/NIE↓, do local cuts/gates incur less performance degradation than turning off Top-K heads or size-matched random cuts?
- Optional: Does the NIE vs ΔPPL plane reveal high-NIE/low-cost mediators missed by head-only ranking; do results transfer across bias types/models?

3. Datasets & Interventions

- Templates: Professions and WinoBias/WinoGender-style probes (gendered pronoun resolution).
- Interventions: swap-gender (he↔she, his↔her) and set-gender/occupation swap (e.g., nurse↔man), with tokenization matched.
- Target step: pronoun prediction step (next-token LM).
- Split: small train/val for ranking/selection; held-out test for final curves.

4. Metrics (Bias & Performance)

- Bias (CMA): TE and NIE_z on the target token.
- Performance (primary): ΔPerplexity (ΔPPL) on WikiText-2 (ΔNLL interchangeable).
- Secondary: Winograd-style Accuracy / −ΔAcc on a small coreference probe.

- Structural cost: #cuts (hard) and ∑|1−α| (soft edit magnitude).
- Selection figure: NIE vs ΔPPL scatter (single-site mediator replacement) to identify high-NIE/low-cost targets.
- Bias–Perplexity Pareto. X-axis = TE/NIE remaining (% of baseline, a.k.a. bias), Y-axis = ΔPerplexity (lower is better). Points closer to the lower-left achieve lower bias at lower performance cost.

5. Methods & Pipeline

- Stage A — CMA localization. On GPT-2 small: compute TE and per-mediator NIE at attn-out / mlp-out on the target step; rank and keep Top-K mediators; plot NIE vs ΔPPL (dataset-level performance change under single-site replacement), plus a fake-replacement control.
- Stage B — Local edits (SFC-lite). Around selected mediators, validate local paths and perform:
    - Local Cut: zero a few high-score connections/sub-dims.
    - Local Gate: multiply selected connections by $\alpha \in [0, 1]$ (reversible; minimize bias with small $\sum|1 - \alpha|$).
- Baselines: Head-off, Random-cut (size-matched), optional Layer-sparsify.

6. Evaluation

- Matched debiasing: At fixed remaining bias thresholds (TE/NIE ≤ τ, % of baseline), compare ΔPPL (and −ΔAcc); plot the Bias–Perplexity Pareto (lower-left is better).
- Matched utility (optional): At fixed ΔPPL budgets (ΔPPL ≤ π), compare achieved remaining bias (TE/NIE % of baseline).

7. Expected Results

- We expect local circuit edits (SFC-lite) to achieve lower ΔPPL at the same remaining bias threshold than head ablation or random cuts.
- On the Bias–Perplexity Pareto frontier, our method should place points closer to the lower-left (lower bias, lower ΔPPL).
- NIE–ΔPPL scatter is expected to surface high-impact / low-cost mediators that head-level ranking misses.
- (Optional) Results generalizes across bias types and remain robust on related models (e.g., DistilGPT2).

8. Milestones & Feasibility

- W1–2: Build probes & interventions; run CMA on GPT-2 small → NIE heatmaps + NIE vs ΔPPL scatter; select Top-K mediators.
- W3–4: Implement Local Cut/Gate + baselines (Head-off/Random) → first Bias–Perplexity Pareto.
- W5–6: Add SAE on selected layers (feature-level SFC-lite), run feature edits + faithfulness checks; refine Pareto.
- W7–8: Robustness (second bias or DistilGPT2 / larger model), ablations & error bars, finalize figures + write-up.