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 node level influence within this set and perform fine-grained edits (subspace edits or soft gating) — SFC-lite (cf. Marks et al., 2025, “Sparse Feature Circuits”) at the node-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 — **Soft local gating vs. discrete ablations.** If we scale mediator signals with continuous, local gates (α∈[0,1]) instead of hard zeroing, do we obtain a better bias–perplexity balance across the full gating curve, compared with head-off and size-matched random cuts?
- RQ2 — **Locate mediators that optimize the trade-off**. Can we locate specific mediators or small mediator sets that achieve superior bias–perplexity trade-offs rather than those selected only for maximum debiasing?
- 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 mediators and keep the Top-K; We then plot NIE vs ΔPPL where single-site replacement means: at one mediator and one intervention location, we replace that mediator's activation with a counterfactual/control activation and re-evaluate the model to obtain its NIE (target token) and ΔPerplexity (dataset-level). We also include a fake-replacement control, applying the same procedure to CMA-irrelevant mediators (low-NIE) and/or non-critical positions to estimate background variance and rule out generic perturbation artifacts.
- Stage B — Local Edits (SFC-lite). Around selected mediators, validate the role of each node and perform:
    - Local Cut: zero out high-impact node sub-dimensions within the activation space, thereby selectively removing their contribution to the target variable.
    - Local Gate: scale the selected nodes' activations by $\alpha \in [0, 1]$ in a reversible manner, suppressing their causal influence, with the regularization of minimizing $\sum|1 - \alpha|$.
- 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).
- Pareto plot construction (greedy; Vig et al., 2020): Rank mediators (or subspace directions) by importance, then greedily add them one at a time; after each addition, recompute remaining bias (TE/NIE % of baseline) and ΔPerplexity to yield a curve whose lower envelope is the Pareto frontier. For local gates, sweep a small grid of α (e.g., 0.25/0.5/0.75/1.0) per step and keep the best point for fair comparison.

7. Expected Results

- **Trade-off improvement.** At matched remaining bias, local cuts / continuous gates should incur smaller ΔPerplexity than head ablation or size-matched random cuts, producing a more favorable Pareto frontier.
- **Locating Pareto-efficient mediators.**Causal localization plus soft gating should locate mediators that sit on/near the Pareto frontier(i.e., low remaining bias with low ΔPPL)and these mediators would likely be missed if we optimize only for debiasing.

- **Optional.** The NIE–ΔPPL plane will surface high-impact/low-cost mediators that head-level ranking may miss, and limited robustness checks (e.g., a second bias or DistilGPT2) will indicate whether such trade-offs transfer.

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.