WHEN SPEED COSTS ACCURACY: CAUSAL FAITHFULNESS OF COMPONENT ATTRIBUTION FOR GENDER-BIAS LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

There are mechanistic techniques that aim to compute the effect of individual components or groups of components on a certain type of output. Moreover, these methods are widely used to target the components that produce stereotypical behavior. Sometimes we want affordable solutions instead of generic tasks, such as IOI, Greater-Than, for quick results – and for this, we need to understand whether cheap surrogates (e.g., DLA, EAP-IG) actually recover causal importance for bias specifically, and how much of this behavior is caused by the indirect effect that is undetectable through these techniques. In our work, we benchmark four attribution methods: DLA, AtP, EAP-IG, DE against activation patching ground truth for gender-bias localization across three models: GPT-2 XL, Llama-3.2-1B, Gemma-2-2B under two ablation baselines on a gender-focused StereoSet variant. We found that the frozen-norm approximation is nearly free – it accounts for only under a sixth of the error, and removing it leaves every ranking metric unchanged. The indirect effects account for the whole problem – 88-110% of the total – leaving direct attribution comparable to or worse than a constant-zero predictor, and recovering only about half of patching's ten most important components. A single backward pass recovers the rest: attribution patching finds 7-9 of that top ten at three forward-equivalent passes rather than 234-1248. The difference does not survive the intervention: ablating each method's selected components at matched capability cost removes a comparable share of the measured bias regardless of which method chose them, so ranking components by individual effect does not determine how good they are as a set. Faithful single-component attribution is therefore necessary but not sufficient for localization: choosing a set of components needs a set-level objective, which none of these methods provide. We report stereotype score and language-modeling scores alongside a continuous bias measure, since a mean logit difference can be driven to zero by shrinking every logit.

1 INTRODUCTION

In recent times, Large Language Models (LLMs) have showcased remarkable performance across diverse natural language processing (NLP) tasks. Still, they also learn several biases from the large datasets on which they are trained [Saeedi et al. (2025); Bai et al. (2025)]. Many bias-mitigation methods treat the model as a black box: they modify the inputs (CDA, [Zmigrod et al. (2019)]), the training objective (adversarial training, INLP, Ravfogel et al., 2020), or the entire parameter set (SFT, RLHF, Lambert et al., 2025).

Recent advancements in *mechanistic interpretability* (MI) have provided methods that allow people to identify the components underlying desired behavior and then edit them. For instance, techniques such as causal mediation analysis – also called Activation Patching – have been used to identify neurons and attention heads associated with gender bias [Vig et al. (2020)]. Other studies have analyzed specific layers to understand how social biases are processed and stored [Prakash et al. (2023); Cai et al. (2024)].

Activation Patching [Vig et al. (2020); Heimersheim & Nanda (2024)] appears costly because it requires one forward pass per component. That is why cheaper alternatives were introduced, such

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

as Attribution Patching [Nanda (2023); Kramár et al. (2024)], Direct Effect, Direct Logit Attribution [Elhage et al. (2021); Wang et al. (2022)], and Edge Attribution Patching [Syed et al. (2023)]. To understand the differences in the use of these methods for the gender bias localization task, we benchmarked them against an ablation ground truth on a gender StereoSet[1] variant. We decomposed the error into two terms: frozen-norm (DLA→DE) and indirect effect (DE→AP). For three chosen models (GPT-2 XL, Llama-3.2-1B, Gemma-2-2B introduced in §4.1), we report component selection divergence, empirical error budget, frozen-norm term, mitigation sweep, and released harness.

2 Related Work

2.1 Societal Bias in Language Models

In the context of artificial intelligence, **bias** is defined as a systematic, unfair prejudice in favor of or against a particular entity, person, or group [Chadha (2024)]. The harm caused by bias in NLP systems is commonly broken down into two categories: representational harm[2] and allocational harm[3] [Blodgett et al. (2020)]. In our work, we focus on harmful gender-related behavior.

2.2 Bias Localization Techniques and Component Attribution

A notable technique is Counterfactual Data Augmentation (CDA) [Zmigrod et al. (2019)]. The CDA approach applies data duplication and swaps protected attributes (e.g., ”he” for ”she”, ”mom” for ”dad”, etc.) It is supposed to change and balance the statistical correlation between specific social or demographic roles. However, this technique appears to be less effective at the scale of modern LLMs [Ding et al. (2024); Zmigrod et al. (2019)]. Localized editing is the alternative: causal mediation analysis identifies the neurons and heads that mediate a gendered prediction [Vig et al. (2020)], and logit-lens readouts [Nostalgebraist (2020)] attribute a prediction to individual writes into the residual stream. Every such pipeline needs an attribution method, and its choice is rarely validated against the causal quantity it approximates.

Activation patching replaces a component’s activation and reruns the model. Attribution patching replaces those passes with a single gradient, edge attribution patching extends it to edges, and EAP-IG [Hanna et al. (2024)] integrates the gradient along a path to reduce first-order error.

Research questions This leaves the following three questions: 1) Do the cheap surrogates recover the causal importance they stand in for, when the target is gender bias rather than an algorithmic task (§6.2)? 2) Where does their error come from – the frozen normalization they are usually blamed for, or the indirect paths they cannot see (§6.3)? 3) Does a more faithful ranking make a better edit, once the selected components are actually ablated (§7.1)?

3 Background

We benchmark five methods for identifying the circuit carrying gender bias. All of them use the same metric, the same set of components, and the same ablation baseline. Full definitions, including the residual decomposition, the folded normalization and the integrated-gradient estimator, are in Appendix A.1.

Residual stream. Let T be the final position of the input context. All quantities below are read at T, and we drop the position index. The final residual stream $\mathbf{r}$ is the sum of the writes of the component catalog $\mathcal{C}$ plus the embedding and bias terms, where $\mathcal{C}$ contains every attention head and every MLP block, so $|\mathcal{C}| = L(H+1)$. The residual passes through a final LayerNorm or RMSNorm, whose learnable weight we fold into the unembedding W_U, leaving a projection P and

[1] https://github.com/McGill-NLP/bias-bench/tree/main/data/stereoset

[2] Occurs when systems disparage, underrepresent, or memorialize negative stereotypes about certain demographic groups.

[3] Occurs when a system distributes resources or chances unjustly based on learned demographic characteristics.

a scalar rescaling $\sigma(\mathbf{r})$; writes are centered under the same convention, so P drops out of every inner product below.

Each dataset item supplies a stereotypical completion s and an anti-stereotypical one a (§5.1), and our metric is their logit difference $m(\mathbf{r}) = \mathbf{u}^\top N(\mathbf{r})$ with $\mathbf{u} = W_U[:, s] - W_U[:, a]$, positive when the model prefers the stereotype. Every method compares a clean activation $\mathbf{a}_c$ ($\mathbf{z}_{l,h}$ for heads, the block output for MLPs) to a baseline $\tilde{\mathbf{a}}_c$, which is $\mathbf{0}$ under zero ablation and the mean of $\mathbf{a}_c$ at T over the analyzed prompts under mean ablation; $\bar{\mathbf{x}}_c$ denotes the residual write the baseline would produce. Positive scores mean c pushes toward the stereotype, for every method below.

3.1 Direct Attribution: DLA and DE

Direct logit attribution (DLA) fixes the normalization scale at its clean value and reads each component's write onto the effective unembedding direction $\hat{\mathbf{e}} = \mathbf{u}/\sigma(\mathbf{r})$. The direct effect (DE) instead replaces c's write in the final residual stream and recomputes the normalization exactly, without recomputing any downstream layer:

$$\mathrm{DLA}(c) = \big(\mathbf{x}_c - \bar{\mathbf{x}}_c\big)^\top \hat{\mathbf{e}}, \qquad \mathrm{DE}(c) = m(\mathbf{r}) - m\big(\mathbf{r} - \mathbf{x}_c + \bar{\mathbf{x}}_c\big). \tag{1}$$

DE is therefore the *complete* direct effect of c: the most that any method reading only the residual stream's projection onto the unembedding can recover. Both cost one forward pass for the whole catalog, and both are blind by construction to every path through which c acts on downstream components. The gap between them has a closed form.

Proposition 1 (Frozen-norm identity)**.** *Let* $\mathbf{r}_c = \mathbf{r} - \mathbf{x}_c + \bar{\mathbf{x}}_c$ *and* $\rho_c = \sigma(\mathbf{r})/\sigma(\mathbf{r}_c)$*. Then*

$$\mathrm{DE}(c) = (1 - \rho_c)\, m(\mathbf{r}) + \rho_c\, \mathrm{DLA}(c). \tag{2}$$

The identity is exact for both baselines and both normalization types (Appendix A.2) and we verify it to 5×10^{-6} on every model; DLA is its $\rho_c \to 1$ limit. The whole DLA–DE discrepancy is one scalar per component: how much removing it rescales the residual stream.

3.2 Activation Patching (AP) and the Error Decomposition

AP intervenes on the component itself and lets the network respond: we replace $\mathbf{a}_c$ with $\tilde{\mathbf{a}}_c$ at position T, rerun the forward pass, and take $\mathrm{AP}(c) = m(\mathbf{r}) - m(\tilde{\mathbf{r}}^{(c)})$. Because every downstream head and MLP is recomputed, AP captures the direct, normalization and indirect effects; it is the causal quantity the other methods approximate and we use it as ground truth. Following Heimersheim & Nanda (2024) we call this a *knockout*: the baseline is a fixed vector, not a counterfactual prompt. AP costs $|\mathcal{C}|$ forward passes per evaluation unit. Since AP and DE differ only through downstream recomputation, $\mathrm{IE}(c) := \mathrm{AP}(c) - \mathrm{DE}(c)$ is the *indirect effect* of c, and the DLA–AP gap telescopes into two measurable parts:

$$\mathrm{AP}(c) - \mathrm{DLA}(c) = \underbrace{\big[\mathrm{DE}(c) - \mathrm{DLA}(c)\big]}_{\text{frozen norm}} + \underbrace{\mathrm{IE}(c)}_{\text{indirect effect}}. \tag{3}$$

3.3 Gradient Estimators: AtP and EAP-IG

Attribution patching (Nanda, 2023; Kramár et al., 2024) replaces the $|\mathcal{C}|$ forward passes of AP with one first-order Taylor expansion around the clean run:

$$\mathrm{AtP}(c) = \big(\mathbf{a}_c - \tilde{\mathbf{a}}_c\big)^\top \left.\frac{\partial m}{\partial \mathbf{a}_c}\right|_{\text{clean}}. \tag{4}$$

The gradient is taken through the intact network, so it flows through every downstream component and the final normalization: AtP sees all three effects to first order, and what it misses is curvature. One forward and one backward pass score the entire catalog, about three forward-equivalents per unit. EAP-IG (Hanna et al., 2024) averages the same gradient over $M = 10$ points along a path and multiplies by the same activation delta; we evaluate a node-level instance whose path lies in the final-token input embedding (Appendix A.1). Path and delta therefore live in different spaces – the gradient is averaged along an input-space path while $\mathbf{a}_c - \tilde{\mathbf{a}}_c$ is the activation-space ablation step – so nothing forces the estimate to be calibrated to $\mathrm{AP}(c)$. Its cost is about $3M$ forward-equivalents per unit.

4 METHODOLOGY

Per evaluation unit, one forward pass gives the direct scores, one backward pass the gradient scores, and $|\mathcal{C}|$ further forward passes the activation-patching ground truth. Each method is scored against AP within the same unit under both ablation baselines, with the DLA–AP gap split into the two terms of Eq. equation 3. The same scores then drive the edit: top-k components ablated jointly, evaluated on held-out items behind a capability gate (§7.1).

4.1 MODELS AND ARCHITECTURES

All experiments are conducted on three autoregressive transformer language models. Our selection fell on the 1-to-2 billion-parameter LLMs (larger LLMs would step out of our computational resources). They provide computational possibilities for fine-tuning and a diversity of architectural evolution: **GPT-2 XL** (48 layers, 25 heads, LayerNorm, multi-head attention), **Gemma-2-2B** (26 layers, 8 query / 4 key-value heads, RMSNorm with additional post-attention and post-FFW norms, gated MLP, final-logit soft-cap), **Llama-3.2-1B** (16 layers, 32 query / 8 key-value heads, RMSNorm, SwiGLU).

4.2 ATTRIBUTION PIPELINE

Component catalog. Every method scores the same catalog $\mathcal{C}$: all attention heads and all MLP blocks, at the final position T. That is $|\mathcal{C}| = 1{,}248$ for GPT-2 XL (48×25 heads $+48$ MLPs), 528 for Llama-3.2-1B ($16 \times 32 + 16$) and 234 for Gemma-2-2B ($26 \times 8 + 26$). Head writes are reconstructed as $\mathbf{z}_{l,h} W_O^{(l,h)}$ from the cached attention-weighted values; MLP writes are read from the block output. Both are verified per layer against the residual-stream difference $\mathbf{r}_l^{\text{post}} - \mathbf{r}_l^{\text{pre}}$ before any score is computed.

The catalog does not contain the token embedding or the output biases. Their share of the clean metric is fixed and cannot be selected by any method. In our work, we report it as the *catalog share* $\sum_{c \in \mathcal{C}} \mathrm{DLA}(c)/m(\mathbf{r})$under zero ablation. It is lowest for GPT-2 XL, whose layer-0 MLP acts as an extended embedding, and we return to it in §7.2.

One evaluation unit, one pass. An evaluation unit consists of one prompt paired with one (stereotype or anti-stereotype) candidate pair. For each unit, one clean forward pass with caching yields $\mathbf{r}$, $\sigma(\mathbf{r})$, and every component's write, from which $\mathrm{DLA}(c)$ and $\mathrm{DE}(c)$ follow for all c at no further cost. One forward and one backward pass yield $\mathrm{AtP}(c)$ for all c; M of each yield $\mathrm{IG}(c)$. $\mathrm{AP}(c)$ requires $|\mathcal{C}|$ additional forward passes, which we batch in 16 components at a time, with one ablation hook per row.

Architecture-specific corrections. Gemma-2 normalizes the attention block's output before adding it to the residual stream, so $\mathbf{z}_{l,h} W_O^{(l,h)}$ is not a head's write. For a fixed scale the norm is diagonal, so the per-head split survives exactly under a recovered per-layer gain, which we verify against $\mathbf{r}_l^{\text{mid}} - \mathbf{r}_l^{\text{pre}}$ and apply to the mean-ablation baseline as well (Appendix A.3). DLA and DE for Gemma heads are therefore exact; one consequence is that knocking out a head also perturbs that scale, so Gemma's head-side indirect effect contains a renormalization term the other two models do not have. AtP and EAP-IG differentiate through the intact norm and need no correction.

Validation gates. Before any score is computed, each run enforces the per-layer residual decomposition, DLA completeness, the frozen-norm identity of Eq. equation 2, invariance of AP to the patching batch size, and non-collinearity of the gradient estimators with DE, which fails if the backward graph has been cut and the estimator has silently become a direct method. Thresholds and the worst observed value per run are in Appendix A.5.

4.3 Error Decomposition: Frozen-Norm and Indirect-Effect Components

Eq. equation 3 splits the DLA–AP gap into two vectors over the catalog. We measure each with a normalized mean absolute error,

$$\mathrm{NMAE}(\mathbf{a}, \mathbf{b}) = \frac{\sum_c |a_c - b_c|}{\sum_c |\mathrm{AP}(c)|} \times 100\%, \tag{5}$$

with the denominator fixed to AP so that the three quantities below share a scale:

$$\begin{aligned} \text{Source A (frozen norm)} &= \mathrm{NMAE}(\mathrm{DLA}, \mathrm{DE}), \\ \text{Source B (indirect effect)} &= \mathrm{NMAE}(\mathrm{DE}, \mathrm{AP}), \\ \text{Total} &= \mathrm{NMAE}(\mathrm{DLA}, \mathrm{AP}). \end{aligned} \tag{6}$$

Two properties of Eq. equation 5 matter for reading the numbers. A predictor that outputs zero for every component scores exactly 100%, so 100% is the constant-zero reference, not a ceiling. Moreover, because the two sources are norms of the two summands in Eq. equation 3, Total $\leq$ A+B by the triangle inequality, the shortfall measures how much the frozen-norm and indirect-effect errors cancel on individual components.

Eq. equation 5 pools over the record set, so it is dominated by the units with the largest $|\mathrm{AP}|$; we therefore also compute it inside each unit and report the median, and the two agree throughout. A large NMAE can mean wrong magnitudes on a right ranking or a wrong ranking, so we also report NMAE after the L_1-optimal rescaling of a method's scores onto AP's (Appendix A.8); a method whose raw and rescaled values coincide is calibrated. All NMAE values carry a 95% bootstrap interval over resampled *examples*, so components of one prompt are resampled together, and method differences on ranking metrics use a paired bootstrap of the per-unit medians.

4.4 Ablation Regimes: Zero and Mean Ablation

Every result is reported under two baselines. Under *zero ablation*, $\tilde{\mathbf{a}}_c = \mathbf{0}$. Under *mean ablation*, $\tilde{\mathbf{a}}_c$ is the mean of $\mathbf{a}_c$ at the final position over the prompts being analyzed, computed once per model. For Gemma-2-2b heads, the mean passes through the same gain as the clean write (Eq. 11), so that the mean-mode DE subtracts a post-norm quantity from a post-norm residual. The four attribution scores use the same baseline within a regime, so DLA, DE, AtP, and EAP-IG are baseline-matched to the AP they are compared against. Using a zero-baseline DLA against mean-baseline AP, a mismatch that is easy to make, inflates Source A and shifts every ranking, because the missed $\bar{\mathbf{x}}_c^\top \hat{\mathbf{e}}$ term varies by component.

Both regimes are *knockouts*: the replacement comes from a fixed vector, not a counterfactual prompt, which moves the model off distribution [Heimersheim & Nanda (2024)]. We use knockout because it is the quantity every method here estimates, with the same $\tilde{\mathbf{a}}_c$ on both sides, and because it is the edit the mitigation applies. The concern can be tested: if it drove the indirect-effect term, mean ablation would shrink that term substantially relative to zero ablation, and §6.3 shows it does not. The mitigation sweep intervenes at every non-BOS position and computes its mean baseline there. Appendix A.13 gives the reason and the measurement.

5 Data

5.1 Dataset: StereoSet for Gender-Bias Localization

Stereoset is a large-scale benchmark designed to measure the LLM's stereotypical behavior across four bias categories: gender, profession, race, and religion. The raw data consists of **intersentence** (created to measure the model's discourse-level reasoning) and **intrasentence** (created to measure the model's sentence-level reasoning, aka fill-the-blank task).

In this work, we restrict attention to the **gender** bias category of the intrasentence task. After filtering the official `test.json`(771 gender examples) and `dev.json`(255 gender examples) for `bias_type == "gender"`, the working dataset contains examples that probe whether the model assigns a higher probability to the stereotypical choice than to the anti-stereotypical.

5.2 Data Processing and Rephrasing

The original StereoSet intrasentence items place the `BLANK` token often in the middle of the sentence. Since autoregressive language models generate tokens strictly left-to-right and can only be meaningfully probed at the **last** position of the input context, this placement is incompatible with both next-token probability scoring and benchmarked techniques. For instance, DLA extracts hidden states and computes per-component contributions at the final sequence position, so if the `BLANK` appeared mid-sentence, subsequent tokens would already be part of the input, breaking the causal assumption. To address this issue, each StereoSet context is **rephrased** using the GPT-o4-mini-high agent, so that the `BLANK` marker is put at the very end of the sentence without context loss. The rephrased item keeps the original item's ID, candidate targets, and sentence meaning. This provides structural compatibility between the benchmark and the autoregressive evaluation pipeline used in this work. For more detailed pipeline – Appendix A.14. Final data stays at repo[4].

6 Experiments

6.1 Experimental Setup

Units. Each item of the rephrased StereoSet gender development split (§5.1) yields one evaluation unit when both candidates are single tokens under the model's tokenizer. Multi-token candidates are skipped rather than truncated. This leaves $n = 75$, 77 and 86 units for GPT-2 XL, Llama-3.2-1B and Gemma-2-2B.

Metrics. Each method is compared with AP within each unit by Spearman ρ and Kendall τ_b (global rank agreement), rank-biased overlap (RBO, $p = 0.9$) and top-10 Jaccard (J10, weighting the top of the ranking), and the Matthews correlation of the sign, which asks whether a method gets the *direction* of a component's effect right – the question a debiasing edit turns on. Top-k sets are by absolute value, since a component can matter in either direction. Rank statistics are medians over units with a bootstrap interval. Definitions and the sign base rates are in Appendix A.8. A ranking drawn uniformly at random per unit gives the floor.

Cost. We count forward-equivalent passes per unit, with a backward pass counted as two forwards. DLA and DE cost 1; AtP 3; EAP-IG $3M$ with $M = 10$; AP $|\mathcal{C}|$, i.e. 234–1,248.

6.2 Ranking Agreement Between Attribution Methods

The ordering is AtP $\gg$ EAP-IG $\gg$ DLA $\approx$ DE $\gg$ Random. AtP reaches median $\rho = 0.887$–0.991, MCC 0.852–0.985 and J10 0.538–0.818, i.e. seven to nine of AP's ten most important components. EAP-IG reaches $\rho = 0.646$–0.816 and J10 0.250–0.538. DLA reaches $\rho = 0.202$–0.376, MCC 0.139–0.309, and J10 0.250–0.333: four or five of the top ten, and the sign of the effect is right a little better than chance. The random floor sits at $\rho = -0.012$ to 0.005 and J10 $= 0.000$. No interval of one method overlaps that of its neighbor in the ordering.

DE does not improve on DLA. Removing the frozen-norm approximation leaves every ranking metric unchanged to the second decimal. On GPT-2 XL under zero ablation, DLA and DE score $\rho = 0.339$ vs. 0.346, identical J10 (0.250) and identical MCC (0.290); the pattern repeats in the other five cells, and on Gemma-2-2B DE is marginally *worse*. The normalization is not what is wrong with direct attribution.

Mean ablation helps every method a little. Every method scores higher under the mean than under zero ablation, by 0.01–0.06 in ρ. The ordering and the gaps between methods do not change. The two regimes are therefore reported side by side throughout rather than choosing one.

[4] `https://anonymous.4open.science/r/bias-mitigation-interpretability-8620/datasets`

6.3 Indirect Effect Analysis

Figure 6a gives the error budget of Eq. equation 6 (full table in Appendix A.10).

The frozen norm is nearly free. Source A is 4.4–14.5% pooled and 3.5–13.4% per unit. It is the largest for GPT-2 XL, the only LayerNorm model, and the smallest for Llama-3.2-1B. It is exactly the term Proposition 1 predicts: per-component, it is $(1 - \rho_c)\,[m(\mathbf{r}) - \mathrm{DLA}(c)]$, so it is large only where removing a component moves the residual norm, which at the final position of a 1–2B model is rare. Its ranking consequence is the DLA $\approx$ DE result of §6.2: none.

The indirect effect is the whole problem. Source B is 88.3–108.8% pooled and 88.9–107.5% per unit. That is the error of a predictor that outputs zero everywhere, or worse. The direct-effect vector is not a noisy estimate of the causal effect; on most components, it is uncorrelated with it. The total DLA–AP error, 88.5–106.6%, is within a few points of Source B alone. Adding the two sources over-counts slightly (A + B $>$ Total by 2–15 points), so the two error terms partially cancel on the individual components, but the cancellation is small.

Off-distribution effects do not explain it. If zero ablation inflated Source B by driving the model off-distribution, mean ablation should reduce it substantially. It reduces it by 5–10 points ($94.1 \rightarrow 88.3$ on GPT-2 XL, $103.5 \rightarrow 97.7$ on Llama, $108.8 \rightarrow 98.5$ on Gemma) and leaves it near 100%. The indirect effect is a property of how the bias is computed, not of the intervention used to measure it.

One backward pass recovers most of it. AtP sees indirect paths to first order, and its NMAE against AP is 14.8–45.1%, a factor of two to six below DE in the same cell. The remainder is curvature. Appendix A.9 verifies that AtP behaves as a Taylor expansion should: its error grows with $|\rho_c - 1|$, the component's share of the residual norm, and the AtP–AP scatter is a line through the origin that fans out at the extremes (Figure 4). The residual AtP error is largest on Gemma-2-2B, whose post-attention norm adds the renormalization term of Appendix A.3 to every head's indirect effect.

The indirect effect is not confined to any depth or component type. Appendix A.11 breaks NMAE against AP down by layer and by heads versus MLPs. For DLA and DE the error sits near 100% at every depth on every model, and is the same for heads and MLPs: a direct method is not right in late layers and wrong in early ones, it is uninformative everywhere. AtP stays low at every depth. The one structured failure is EAP-IG on GPT-2 XL, whose error is concentrated in layer 0 – the extended embedding – where its input-space integration path is furthest from the activation-space ablation it is meant to estimate.

6.4 Methods Select Different Components, and None Beats the Ground Truth's Own Stability

The methods disagree about *where* the bias sits (Figure 1). On GPT-2 XL, AP's top ten is enriched $1.45\times$ in the deepest fifth of the network and still draws from the shallowest fifth at $0.53\times$ base rate; DLA and DE over-concentrate the deep end ($2.28\times$, $2.40\times$) and almost never nominate the first two fifths ($0.05\times$, $0.13\times$), seeing only what writes directly to the unembedding, while EAP-IG is enriched $3.04\times$ in the shallowest fifth, the layer-0 artifact of §6.3. AtP tracks AP's profile most closely on all three models. The disagreement is structural, not noise around a shared ordering.

A second question is how reproducible a method's aggregated top-10 set is across resamples of the units, against its overlap with AP's (Appendix A.7). AP's own self-agreement – the ceiling any method could reach – is J10 $= 0.43$ on GPT-2 XL and 0.82 on the other two, and AtP's agreement with AP is 0.43, 0.82, 0.82: exactly that ceiling. DLA reaches 0.33, 0.67, 0.82 and EAP-IG 0.25, 0.67, 0.82; on GPT-2 XL DLA is *more* self-consistent (0.67) than AtP (0.33), a stable answer to a different question. Sub-sampling shows every metric converged well below our n, so n is set by the cost of AP, not estimator variance. Two consequences carry into the next section: at small k a faithful ranker's population top-k *is* AP's, and the attainable ceiling is the reproducibility of the target itself.

Table 1: Bias reduction (%) on held-out items at a matched capability budget: the largest reduction any k achieves within that perplexity increase, with 95% intervals over held-out items (Random: min–max over five permutations). AOR is the area between a method's curve and Random's on one interval per model. Configurations above +25% perplexity are excluded; a value repeated across budgets means the method's curve ends at the gate. The permutation floor reaches 0.6–2.8% at 5% cost and 0.6–4.6% at 25% (AOR 0.000 by construction) and is omitted from the rows.

Model	Method	at 5%	at 10%	at 25%	AOR
Llama-3.2-1B	DLA	13.6 [10.2, 17.4]	13.6	13.6	0.095
	AtP	26.4 [21.5, 32.8]	29.1	39.9	0.238
	EAP-IG	26.4 [21.5, 32.8]	26.4	26.4	0.202
	AP (skyline)	16.9 [13.4, 21.0]	16.9	16.9	0.125
GPT-2 XL	DLA	27.3 [20.8, 35.0]	34.2	34.2	0.240
	AtP	27.8 [21.9, 35.8]	34.3	34.3	0.258
	EAP-IG	29.0 [16.5, 36.4]	29.0	35.9	0.243
	AP (skyline)	28.5 [22.2, 36.7]	35.0	35.0	0.266
Gemma-2-2B	DLA	13.1 [10.7, 16.9]	23.9	36.9	0.222
	AtP	14.9 [12.4, 18.5]	22.6	39.4	0.200
	EAP-IG	14.9 [12.4, 18.5]	22.6	39.4	0.200
	AP (skyline)	13.1 [10.5, 16.6]	24.8	39.7	0.207

7 RESULTS

7.1 FAITHFULNESS DOES NOT DETERMINE MITIGATION

Scores are aggregated into one population ranking by mean rank across units. We take the signed top-k and mean-ablate those components jointly at every non-BOS position, $k \leq 32$. Selection uses the development file and every number below is on the held-out file: bias is the mean stereotype minus anti-stereotype logit difference there, capability cost the relative perplexity increase on generic text, and configurations above +25% perplexity are collapsed models and excluded. AP is a *skyline*, selecting with exact ground truth, and five random permutations give the floor. Because cost is not monotone in k, we report at each budget the largest reduction any k achieves within it.

On GPT-2 XL the four methods remove 27–29% of the measured bias at 5% capability cost and 29–35% at 10%, every interval overlapping every other, against 0.6% for Random (Table 1, Figure 7); on Gemma-2-2B it is 13–15% at 5% and 37–40% at 25%. The AP skyline does not dominate – it leads GPT-2 XL by 1.2 points inside a 14-point interval and trails AtP on Gemma-2-2B – and area over random is within 0.03 across the four methods on both.

On Llama-3.2-1B, DLA reaches 13.6% and the AP skyline 16.9% before the gate while AtP reaches 39.9%. This cannot be the intrinsic ordering, since AP is the ground truth; it is the cost of the first MLP each method nominates – +39% perplexity for DLA at $k = 4$, +52% for AP at $k = 6$, +45% for EAP-IG at $k = 8$ – while AtP's top 16 are all heads. Restricting the pool to heads (Appendix A.13) removes the discontinuity: at $k = 8$ the methods remove 29–33% on Llama-3.2-1B, 11–14% on GPT-2 XL and 23–25% on Gemma-2-2B at comparable cost. EAP-IG's collapse is the fingerprint made concrete: at $k = 12$ it adds MLP 0 and perplexity rises by +830%.

At small k the gradient methods and AP select the *same* components: the population top-k Jaccard between AP and AtP is 1.00 on GPT-2 XL ($k \leq 8$) and Llama-3.2-1B ($k = 4$), and 0.78–0.88 at $k = 16$, so their equal effect is identity rather than coincidence. DLA's overlap with AP is 0.14–0.78: it selects a different set and removes a comparable share anyway. Single-component causal importance under knockout does not predict the joint effect of a set, because components compensate and the bias is not carried by a few of them. StereoSet's SS and LMS (Appendix A.13) tell the same story.

7.2 CROSS-ARCHITECTURE COMPARISON

The frozen-norm term follows the normalization. It is largest on GPT-2 XL (14.5% zero, 7.5% mean), the LayerNorm model, where centering couples every component's write to σ; smallest

on Llama-3.2-1B (5.8%, 4.4%); and intermediate on Gemma-2-2B (12.9%, 11.1%), whose post-attention norm adds a second rescaling per head (Appendix A.3). The same extra term makes Gemma-2-2B the model on which AtP's first-order estimate is least accurate (45.1% NMAE under zero ablation against 18.2% on GPT-2 XL), because knocking out a head there perturbs a normalization scale the gradient linearizes. GPT-2 XL has the smallest catalog share: its layer-0 MLP behaves as an extended embedding, so a larger fraction of the clean metric is carried by writes no method can select, and its layer-0 knockouts dominate the pooled budget (§4.3). None of these differences changes the ordering of the methods, the budget's conclusion, or the mitigation result.

8 Conclusion

We set out to answer three questions about the attribution methods used to localize social bias: whether the cheap surrogates recover the causal importance they stand in for, where their error comes from, and whether a more faithful ranking makes a better edit. We decomposed the DLA–AP gap into a frozen-norm term with a closed-form identity and an indirect-effect term, measured both against activation patching on three architectures and two ablation baselines, then used each ranking as an edit under a held-out evaluation and a capability gate.

The results are that direct attribution does not recover causal importance for gender bias: its error is 88–110% of $\sum|\mathrm{AP}|$, no better than a constant-zero predictor, and it recovers four or five of patching's ten most important components with the sign right little better than chance. That error is not the normalization, the term direct methods are usually blamed for: the frozen-norm term is 4–14%, exactly what Proposition 1 predicts, and removing it changes no ranking. It is the indirect effect, structurally invisible to any method reading only the residual stream's projection onto the unembedding and present at every depth and for both component types. One backward pass recovers most of it: AtP reaches $\rho = 0.891$–0.991 at three forward-equivalent passes, is calibrated to within 7% in scale, and agrees with AP's top ten exactly as often as AP agrees with itself. The practical choice is therefore not cheap versus faithful – AtP is both – and what cheap buys badly is DLA.

None of that transfers to the intervention. At matched capability cost, sets chosen by a method with $\rho = 0.2$ and by the ground truth itself remove comparable shares of bias on two of three models. On the third the ordering is set by which method first nominates an MLP, not by how well it ranks. At small k the faithful methods and AP select the same components, so their equal effect is identity, while DLA selects a different set and matches them anyway. Ranking by individual effect is necessary but not sufficient.

Two directions follow. The first is set-level selection – mask optimization or search under an explicit capability constraint – evaluated as here, on held-out data behind a perplexity gate, with the ceiling of §6.4 bounding what any selector can reproduce. The second is to test how far the picture holds: under interchange interventions, across positions, and for other bias types. The released harness runs both against the same ground truth.

Limitations Our ground truth is knockout, not interchange intervention. We chose it because it is what every surrogate estimates and what the mitigation applies, but the two can disagree. Selection is at the final position while the intervention is at all positions, so earlier-position paths are outside every method's scope, and single-component necessity is not circuit membership. We use one dataset, one language and one bias type, and the held-out bias measure has 166–192 items. The rephrasing is LLM-generated and was not independently validated, our EAP-IG integrates in input space so its results hold for that variant, and the mitigation is a fixed ablation, not a trained edit.

AI use & Ethics statement

We used Grammarly[5] and Claude[6] to aid in paraphrasing while writing this work, ensuring that our language is clear and respectful. In this work, we promote the ideals of equality and fairness, highlighting gender biases in LLMs. We take responsibility for analyzing this sensitive topic and contributing to the discourse on fairness and equity.

[5] https://app.grammarly.com

[6] https://claude.ai

REFERENCES

X. Bai et al. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences (PNAS)*, 2025. doi: 10.1073/pnas.2416228122. URL https://www.pnas.org/doi/10.1073/pnas.2416228122.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.485.

Y. Cai et al. Locating and mitigating gender bias in large language models. *arXiv preprint arXiv:2403.14409*, 2024. URL https://arxiv.org/abs/2403.14409.

Kabir Chadha. Bias and fairness in artificial intelligence: Methods and mitigation strategies. *International Journal for Research Publication and Seminar*, 15(3):36–49, 2024. doi: 10.36676/jrps.v15.i3.1425.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using Large Language Models: Data perspectives, learning paradigms and challenges, 2024. URL https://arxiv.org/abs/2403.02990.

N. Elhage, N. Nanda, et al. A mathematical framework for transformer circuits. 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. doi: 10.48550/arXiv.2404.15255.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024. doi: 10.48550/arXiv.2403.00745.

N. Lambert et al. Reinforcement learning from human feedback. *arXiv preprint arXiv:2504.12501*, 2025. URL https://arxiv.org/abs/2504.12501.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models, 2020.

Neel Nanda. Attribution patching: Activation patching at industrial scale. https://www.neelnanda.io/mechanistic-interpretability/attribution-patching, 2023. Blog post.

Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Nostalgebraist. Interpreting gpt: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

N. Prakash et al. Layered bias: Interpreting bias in pretrained large language models. In *BlackboxNLP*, 2023. URL https://aclanthology.org/2023.blackboxnlp-1.22.pdf.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection, 2020. URL https://arxiv.org/abs/2004.07667.

P. Saeedi et al. Heuristics and biases in ai decision-making: Implications for responsible agi. *arXiv preprint arXiv:2410.02820*, 2025. URL https://arxiv.org/abs/2410.02820.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*, 2023.

J. Vig et al. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology, 2019. URL https://arxiv.org/abs/1906.04571.

A Appendix

A.1 Estimator Definitions

Residual stream and normalization. The final residual stream is a sum of writes,

$$\mathbf{r} \;=\; \mathbf{x}_{\mathrm{emb}} \;+\; \sum_{c\in\mathcal{C}} \mathbf{x}_c \;+\; \mathbf{b}, \tag{7}$$

where $\mathbf{x}_{\mathrm{emb}}$ is the token and positional embedding, $\mathbf{b}$ collects the bias terms, a head writes $\mathbf{x}_{l,h} = \mathbf{z}_{l,h} W_O^{(l,h)}$ with $\mathbf{z}_{l,h} \in \mathbb{R}^{d_{\mathrm{head}}}$ its attention-weighted value vector, and an MLP writes its block output $\mathbf{x}_{l,\mathrm{mlp}}$. The residual then passes through the final normalization before the unembedding W_U. Folding the normalization's learnable weight $\boldsymbol{\gamma}$ and bias into W_U (Nanda & Bloom, 2022) leaves

$$N(\mathbf{r}) = \frac{P\mathbf{r}}{\sigma(\mathbf{r})}, \qquad \sigma(\mathbf{r}) = \sqrt{\tfrac{1}{d}\|P\mathbf{r}\|^2 + \epsilon}, \tag{8}$$

with $P = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$ for LayerNorm and $P = I$ for RMSNorm. Folding is a reparametrization: it does not change the model's outputs, and it applies to all three models, each of which has a learnable $\boldsymbol{\gamma}$ in its checkpoint.

Metric.

$$m(\mathbf{r}) = \mathbf{u}^\top N(\mathbf{r}), \qquad \mathbf{u} = W_U[:, s] - W_U[:, a]. \tag{9}$$

The unembedding bias cancels in the difference. For Gemma-2, m is read before the final-logit soft-cap, a monotone per-logit squashing applied after the linear readout, whose effect we report separately.

Completeness. Under zero ablation $\mathrm{DLA}(c) = \mathbf{x}_c^\top \hat{\mathbf{e}}$ and, substituting Eq. equation 7 into Eq. equation 9, $m(\mathbf{r}) = \sum_c \mathrm{DLA}(c) + (\mathbf{x}_{\mathrm{emb}} + \mathbf{b})^\top \hat{\mathbf{e}}$, which we enforce at runtime (Appendix A.5).

Integrated-gradient estimator. Let $\mathbf{x}_T$ be the clean final-token embedding and $\tilde{\mathbf{x}}_T$ its mean over the development split, and for $\alpha \in [0, 1]$ let $\mathbf{g}_c(\alpha)$ denote $\partial m / \partial \mathbf{a}_c$ on the forward pass whose final-token embedding is $\tilde{\mathbf{x}}_T + \alpha(\mathbf{x}_T - \tilde{\mathbf{x}}_T)$. Then

$$\mathrm{IG}(c) = \big(\mathbf{a}_c - \tilde{\mathbf{a}}_c\big)^\top \frac{1}{M} \sum_{j=1}^{M} \mathbf{g}_c(\alpha_j), \qquad \alpha_j = \frac{j - \frac{1}{2}}{M}, \tag{10}$$

with $M = 10$ midpoint steps; a single step at $\alpha = 1$ reduces Eq. equation 10 to AtP (Eq. equation 4).

A.2 Proof of Proposition 1

Write $\mathbf{r}' = \mathbf{r}_c$, and write σ, σ' for $\sigma(\mathbf{r})$, $\sigma(\mathbf{r}')$. By Eq. equation 9 and the linearity of P and $\mathbf{u}^\top$,

$$\begin{aligned} m(\mathbf{r}') &= \frac{\mathbf{u}^\top P\mathbf{r} - \mathbf{u}^\top P(\mathbf{x}_c - \bar{\mathbf{x}}_c)}{\sigma'} = \frac{\sigma}{\sigma'}\,\frac{\mathbf{u}^\top P\mathbf{r}}{\sigma} - \frac{\sigma}{\sigma'}\,\frac{(\mathbf{x}_c - \bar{\mathbf{x}}_c)^\top P\mathbf{u}}{\sigma} \\ &= \rho_c\, m(\mathbf{r}) - \rho_c\, \mathrm{DLA}(c), \end{aligned}$$

using $P^\top = P$. Substituting into Eq. equation 1, $\mathrm{DE}(c) = m(\mathbf{r}) - \rho_c\, m(\mathbf{r}) + \rho_c\, \mathrm{DLA}(c)$. □

The argument never uses the value of $\bar{\mathbf{x}}_c$, so the identity holds for zero and mean ablation alike. The additive ϵ in σ enters only through ρ_c, which is computed on the actual ablated residual.

A.3 Gemma-2 Post-Attention Normalization

Gemma-2 applies an RMSNorm to the attention block's output before adding it to the residual stream, so $\mathbf{z}_{l,h} W_O^{(l,h)}$ is not the head's residual write. For a fixed scale S_l the norm is a diagonal map, $\mathbf{y} = \mathbf{x} \odot (\mathbf{w}_l / S_l)$, so the per-head split survives exactly:

$$\mathbf{x}_{l,h} = \left(\mathbf{z}_{l,h} W_O^{(l,h)}\right) \odot \frac{\mathbf{w}_l}{S_l}, \tag{11}$$

where S_l is the clean-run scale. We recover the gain $\mathbf{w}_l / S_l$ from the cache, verify that Eq. equation 11 reconstructs $\mathbf{r}_l^{\mathrm{mid}} - \mathbf{r}_l^{\mathrm{pre}}$, and apply the same gain to the mean-ablation baseline $\bar{\mathbf{x}}_{l,h}$. Two consequences follow. DLA and DE for Gemma heads are post-gain and therefore exact. And knocking out a head at $\mathbf{z}_{l,h}$ also changes S_l, so for Gemma-2 the head half of the indirect-effect term contains a renormalization component that GPT-2 XL and Llama-3.2-1B do not have. AtP and EAP-IG take gradients with respect to $\mathbf{z}_{l,h}$ through the intact norm, so they need no correction.

A.4 The Integrated-Gradient Variant

We follow the integrated-gradient formulation of attribution patching, but the estimator we evaluate makes two concrete choices that a reader should not have to infer, because they determine what the resulting numbers mean.

First, the *integration path*. We interpolate the last-token input embedding between the clean embedding and a baseline embedding (the mean of `blocks.0.hook_resid_pre` at the final position over the development set), and accumulate gradients with respect to node activations at m midpoint-rule steps along that path. Only the final token's embedding is interpolated; the rest of the sequence is left at its clean value.

Second, the *delta term*. The attribution of a component c is $\langle a_c^{\mathrm{clean}} - a_c^{\mathrm{base}},\ \overline{\nabla_{a_c} \mathcal{M}} \rangle$, where a_c^{base} is the same ablation baseline that activation patching replaces c with (zero, or the component's mean activation), so the delta lives in *activation* space.

These two baselines are not the same object: the path runs from a mean *embedding* while the delta runs from an ablation baseline in *activation* space. The gradient is therefore integrated along a trajectory that does not terminate at the point the delta is measured from, and the estimator is not calibrated against the intervention it approximates. We keep this construction because the delta must match what AP actually does for the comparison against AP to be meaningful, and we report the consequence rather than hiding it: the raw NMAE of this estimator is not a statement about how well integrated gradients can work in principle.

The EAP-IG results are specific to this estimator. Our EAP-IG numbers describe the variant defined in Section 3.3: an integrated gradient along the last-token input-embedding path, with an activation-space delta taken against the ablation baseline. Published variants differ in exactly the dimension that matters here. EAP-IG-inputs interpolates between a clean and a *corrupted* input, so its path endpoints and its delta refer to the same pair and are mutually calibrated; EAP-IG-activations interpolates node activations directly. Our setting has no corrupted counterfactual prompt at all – the ground truth is ablation-based activation patching – so neither published variant transfers unchanged. We therefore claim only that *this* estimator, at $3m$ forward-equivalent passes, does not recover AP's ranking as well as a single-backward-pass AtP does; we do not claim that integrated gradients cannot.

One scale anomaly is unexplained. The L_1-optimal rescaling diagnostic separates ranking error from magnitude error, and for EAP-IG it behaves very differently on GPT-2 XL than on the other two models. On Llama-3.2-1B and Gemma-2-2B the optimal scale is close to one (0.89–0.95), so raw and calibrated NMAE nearly coincide. On GPT-2 XL it is 0.044 (zero ablation) and 0.128 (mean ablation), and calibration moves NMAE from 199.3% to 98.6%: almost all of the raw error

is magnitude, not ranking. The same model also shows the largest displacement of attributed mass toward early layers, $5.7\times$ AP's first-quintile mass against $1.2\times$ for Llama and $2.3\times$ for Gemma.

It is tempting to attribute this to the path/delta baseline mismatch described above, but that mismatch is present in all three models and so cannot by itself explain an anomaly specific to one. GPT-2 XL differs from the other two in ways that plausibly interact with an embedding-space integration path – its LayerNorm subtracts the mean, and its layer-0 MLP behaves as an extension of the token embedding – but we have not isolated which, and we report the anomaly as open rather than resolved.

A.5 Validation Gates

Every run enforces, before scoring:

1. The per-layer residual decomposition $\mathbf{r}_l^{\text{post}} = \mathbf{r}_l^{\text{pre}} + \mathbf{x}_{l,\text{attn}} + \mathbf{x}_{l,\text{mlp}}$ to 10^{-3}.
2. DLA completeness, $\sum_c \text{DLA}(c) + (\mathbf{x}_{\text{emb}} + \mathbf{b})^\top \hat{\mathbf{e}} = m(\mathbf{r})$.
3. The frozen-norm identity, Eq. equation 2, to 10^{-4} in both ablation modes.
4. AP invariance to the patch batch size.
5. For the gradient estimators, non-collinearity with DE ($\text{corr}(\text{IG}, \text{DE}) < 0.995$), which fails if the backward graph has been cut and the estimator has silently become a direct method.

A further control, reported but not gated: the last-layer MLP writes directly into $\mathbf{r}$ with nothing downstream of it at position T, so its indirect effect is zero and $\text{AP} = \text{DE}$ there exactly.

A.6 Cross-Architecture Detail

GPT-2 XL has the smallest catalog share: its layer-0 MLP acts as an extended embedding, so more of the clean metric is carried by writes no method can select, and its layer-0 knockouts dominate the pooled budget. Gemma-2-2B's post-attention norm (Appendix A.3) adds a renormalization term to every head's indirect effect, which is why AtP's first-order estimate is least accurate there: knocking out a head perturbs a scale the gradient linearizes.

A.7 Selection Divergence and Stability

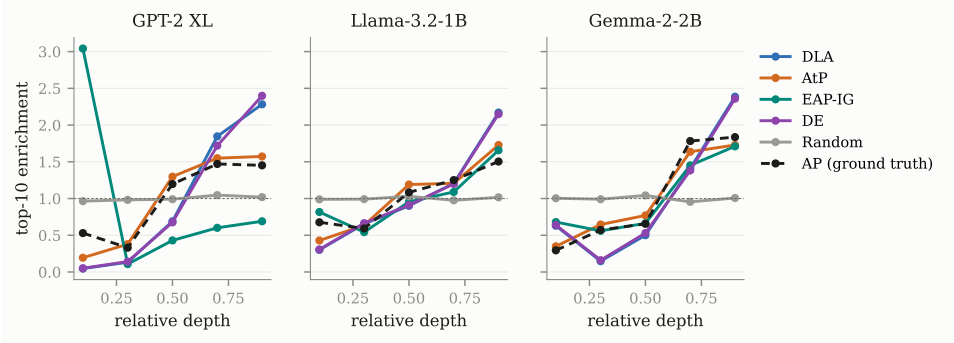


Figure 1: Enrichment of each depth quintile in a method's top-10 set relative to a uniform draw (1.0).

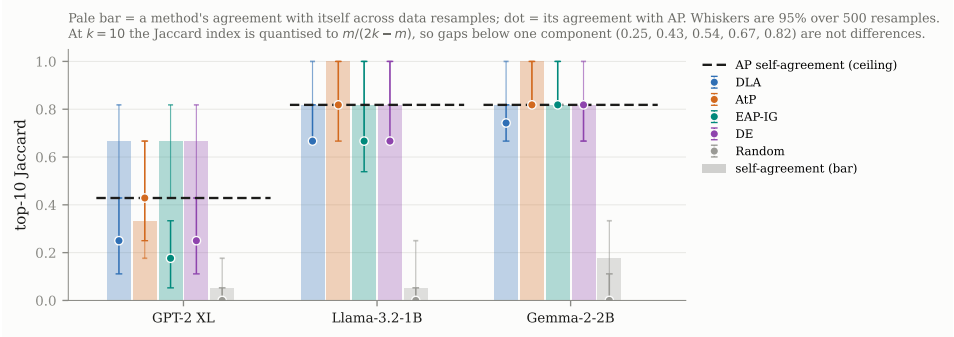


Figure 2: Pale bar: a method's top-10 self-agreement across resamples of the evaluation units; dot: its agreement with AP's top-10; dashed: AP's own self-agreement, the ceiling any method could reach. At $k = 10$ the Jaccard index is quantized $(0.25, 0.43, 0.54, 0.67, 0.82)$, so gaps below one step are not differences.

A.8 Metric Definitions and Scale Diagnostics

Definitions. Spearman ρ and Kendall τ_b measure global rank agreement. Rank-biased overlap (RBO, $p = 0.9$) weights the top of the ranking. Top-10 Jaccard (J10) is the overlap of the two methods' ten largest-$|\cdot|$ components. The Matthews correlation of sign (MCC) treats each component's sign as a binary label. We also report raw sign agreement against its base rate, the accuracy of a predictor that always guesses the majority sign. Rank statistics are medians over units with a bootstrap interval, and top-k sets are taken with respect to absolute value.

Pooled and per-unit NMAE. Eq. equation 5 sums over the whole record set, so it is dominated by the few units with the largest $|\mathrm{AP}|$; on GPT-2 XL these are layer-0 MLP knockouts with $|\mathrm{AP}| \approx 10$ against a typical 0.1. We therefore also compute Eq. equation 5 inside each evaluation unit and report the median.

Scale versus shape. To separate a wrong magnitude from a wrong ranking we find the scalar $s^\star$ minimizing $\sum_c |s\,a_c - \mathrm{AP}(c)|$, the L_1-optimal rescaling (an $|a|$-weighted median of $\mathrm{AP}(c)/a_c$), and report $\mathrm{NMAE}_{\mathrm{cal}}(\mathbf{a}) = \mathrm{NMAE}(s^\star\mathbf{a}, \mathrm{AP})$.

Uncertainty. All NMAE values carry a 95% percentile bootstrap interval from $1{,}000$ resamples of *examples* (not records), so components of the same prompt are resampled together. Differences between methods on ranking metrics use a paired bootstrap of the per-unit medians.

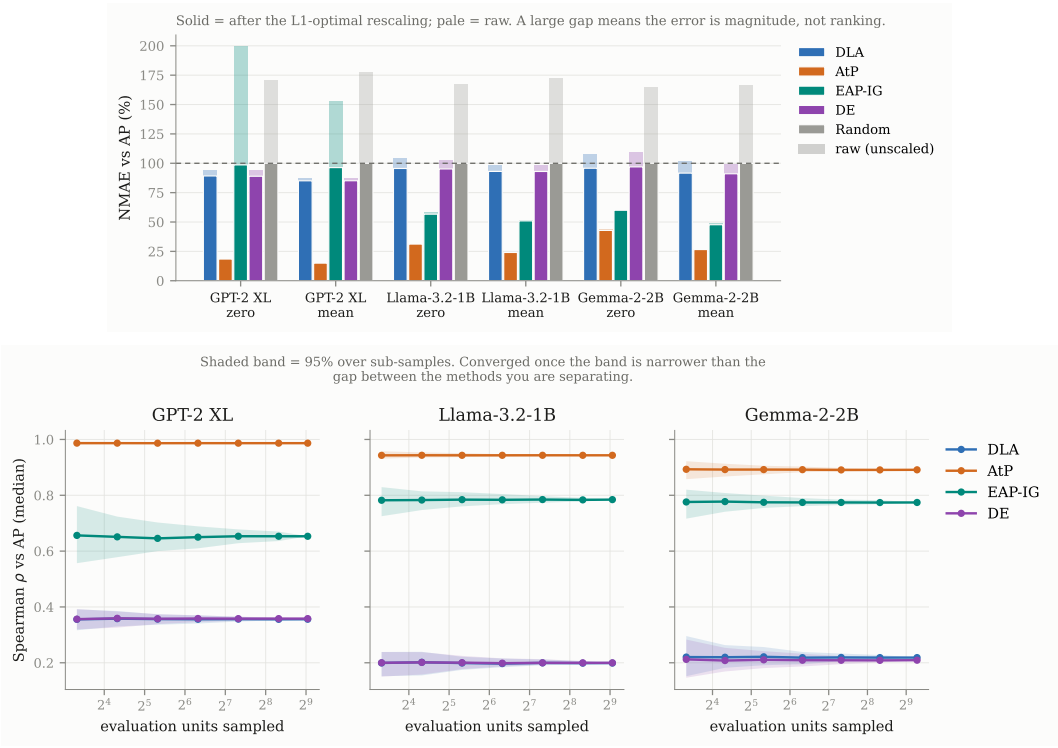


Figure 3: Top: raw versus L_1-calibrated NMAE. AtP's coincide ($s^\star \in [0.99, 1.07]$); EAP-IG's on GPT-2 XL falls from 199% to 99% with $s^\star = 0.04$, so its error there is scale, but the ranking it leaves behind is still poor. Bottom: median Spearman against AP as a function of the number of evaluation units, with 95% bands over sub-samples; every method has converged well below the n we use.

A.9 AtP Verification

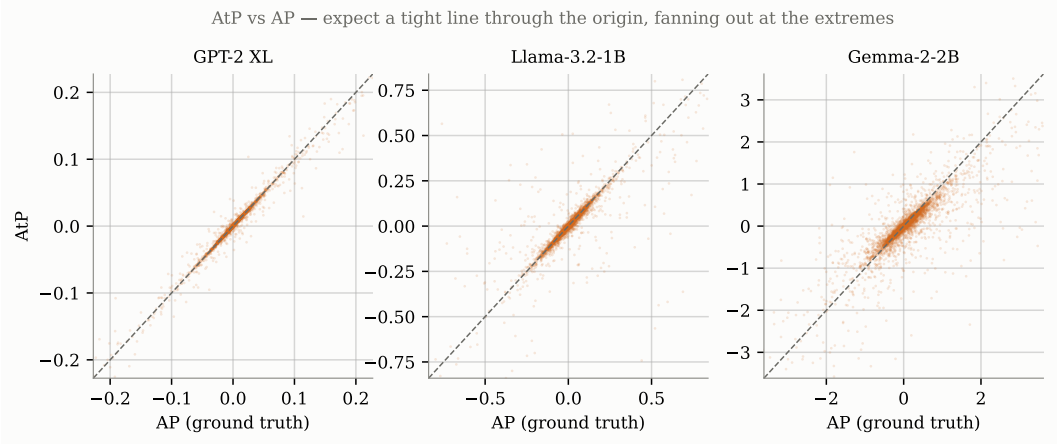


Figure 4: AtP against AP per component. A first-order estimate should give a tight line through the origin that fans out where the intervention is large; a kink or a second branch would indicate a bug.

Table 2: Cheap attribution methods against activation-patching ground truth. NMAE is normalised by $\sum|\mathrm{AP}|$; *cal.* is NMAE after the L_1-optimal rescaling, so a large raw–cal. gap means the error is magnitude rather than ranking. ρ, τ_b, RBO and J10 are medians over evaluation units (one per example–candidate pair). Sign agreement is reported against its empirical base rate. n is the number of evaluation units; *cost* is forward-equivalent passes per unit (a backward pass counts as two forwards, m is the number of integration steps), against one forward per component for AP.

Model	Mode	n	Method	NMAE	cal.	ρ	τ_b	RBO	J10	MCC	cost
GPT-2 XL	zero	527	DLA	94.2	89.2	0.356	0.273	0.367	0.250	0.295	1
			AtP	18.5	18.5	0.987	0.961	0.723	0.818	0.971	3
			EAP-IG	199.3	98.6	0.653	0.542	0.409	0.250	0.639	3m
			DE	94.0	89.0	0.358	0.274	0.380	0.333	0.294	1
			Random	170.5	100.0	0.001	0.000	0.002	0.000	0.000	0
GPT-2 XL	mean	527	DLA	87.7	85.1	0.375	0.295	0.423	0.333	0.313	1
			AtP	15.0	15.0	0.991	0.977	0.731	0.818	0.984	3
			EAP-IG	152.5	96.3	0.684	0.576	0.557	0.429	0.651	3m
			DE	87.7	85.1	0.373	0.294	0.427	0.333	0.311	1
			Random	177.2	100.0	0.002	0.002	0.002	0.000	0.002	0
Llama-3.2-1B	zero	533	DLA	104.4	95.6	0.199	0.149	0.444	0.333	0.147	1
			AtP	31.3	31.2	0.943	0.893	0.648	0.538	0.919	3
			EAP-IG	58.0	56.7	0.784	0.653	0.631	0.538	0.650	3m
			DE	102.7	95.2	0.200	0.149	0.468	0.333	0.148	1
			Random	167.1	100.0	0.002	0.001	0.010	0.000	0.002	0
Llama-3.2-1B	mean	533	DLA	98.7	93.1	0.224	0.171	0.476	0.333	0.169	1
			AtP	24.1	24.1	0.959	0.925	0.760	0.667	0.947	3
			EAP-IG	51.8	51.0	0.791	0.669	0.713	0.538	0.667	3m
			DE	98.0	93.0	0.224	0.170	0.484	0.333	0.168	1
			Random	172.6	100.0	-0.001	-0.000	0.009	0.000	0.002	0
Gemma-2-2B	zero	622	DLA	107.8	95.8	0.219	0.159	0.378	0.250	0.138	1
			AtP	43.5	42.8	0.891	0.814	0.609	0.538	0.857	3
			EAP-IG	60.2	60.0	0.774	0.647	0.585	0.429	0.662	3m
			DE	109.7	96.9	0.209	0.154	0.367	0.250	0.137	1
			Random	164.5	100.0	-0.008	-0.006	0.029	0.000	-0.004	0
Gemma-2-2B	mean	622	DLA	101.4	91.6	0.279	0.210	0.453	0.333	0.194	1
			AtP	26.5	26.5	0.942	0.878	0.793	0.667	0.908	3
			EAP-IG	48.9	47.7	0.822	0.700	0.735	0.538	0.698	3m
			DE	99.4	91.1	0.284	0.214	0.469	0.333	0.197	1
			Random	166.8	100.0	0.002	0.000	0.029	0.000	0.001	0

Table 3: Error budget. Source A is the frozen-norm term $\mathrm{NMAE}(\mathrm{DLA}, \mathrm{DE})$ and Source B the indirect-effect term $\mathrm{NMAE}(\mathrm{DE}, \mathrm{AP})$, both normalised by $\sum|\mathrm{AP}|$. *pooled* uses one global denominator; *per-unit* is the median of the per-unit value and is the robust companion.

		Source A		Source B		Total	
Model	Mode	pooled	per-unit	pooled	per-unit	pooled	per-unit
GPT-2 XL	zero	13.6	11.4	94.0	93.3	94.2	93.3
GPT-2 XL	mean	6.6	5.6	87.7	87.9	87.7	87.6
Llama-3.2-1B	zero	5.1	4.3	102.7	103.4	104.4	104.7
Llama-3.2-1B	mean	3.9	3.3	98.0	98.1	98.7	99.0
Gemma-2-2B	zero	11.7	10.2	109.7	108.4	107.8	106.5
Gemma-2-2B	mean	9.9	8.4	99.4	98.7	101.4	100.6

A.10 Full Intrinsic Tables

A.11 Error by Depth and Component Type

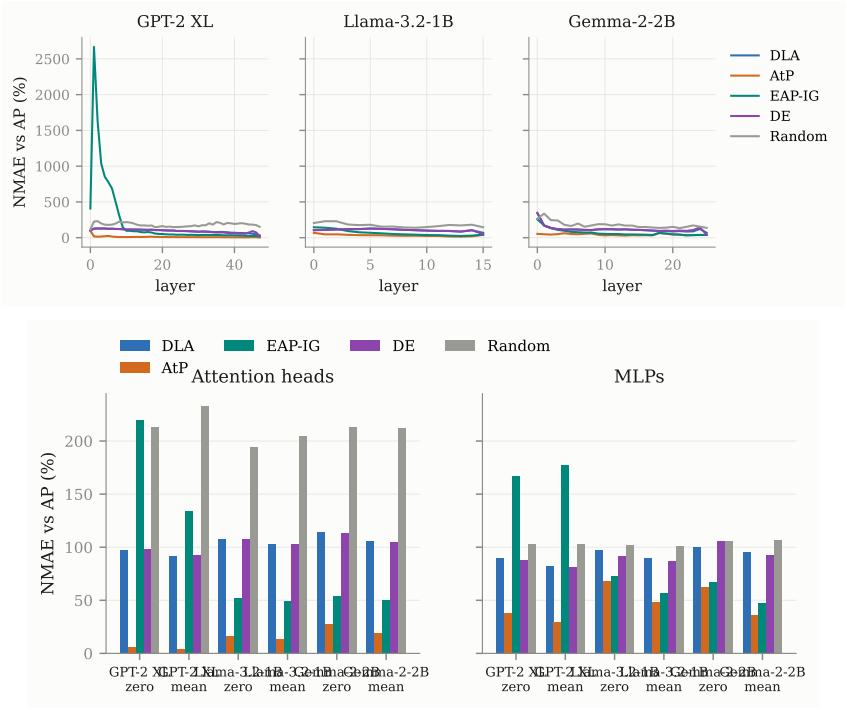


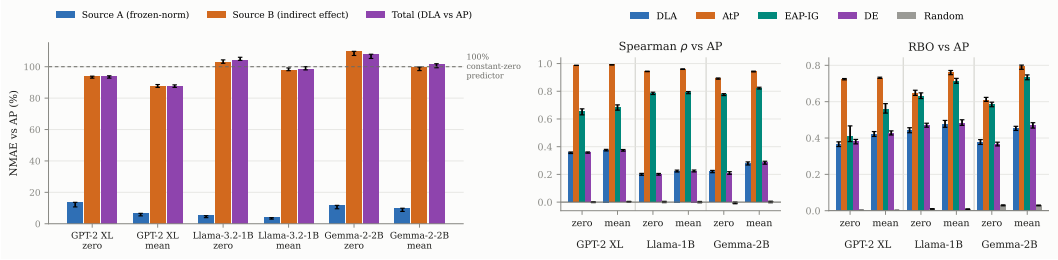
Figure 5: NMAE against AP by layer (top) and by component type (bottom). DLA and DE sit near the constant-zero reference at every depth and for both types; EAP-IG's error on GPT-2 XL is concentrated in layer 0.

A.12 Error budget

A.13 Mitigation: Full Sweep and Robustness

The mean baseline for the mitigation edit is computed over all non-BOS positions of development prompts and generic text. The BOS position is excluded from the mean and from the intervention because early-layer activations there act as an attention sink: replacing them degrades the model independently of any bias-relevant computation. Restricting the selection pool to attention heads (Table 5) removes the discontinuity that dominates the mixed sweep, where the cost of a configuration is set by whether the method happened to nominate an MLP.

Tables 6, 8, 7 list every row of the sweep, gate included and excluded: the number of heads and MLPs each method's top-k contains, the share of non-embedding parameters they account for, StereoSet's SS, LMS and ICAT, and the perplexity cost.



(a) Error budget (b) Rank agreement with AP

Figure 6: **(a)** NMAE of DLA against DE (frozen norm), DE against AP (indirect effect) and DLA against AP (total), normalized by $\sum|\mathrm{AP}|$; the dashed line is the constant-zero predictor, not a ceiling. **(b)** Median per-unit Spearman ρ and RBO against AP with bootstrap intervals. DE tracks DLA everywhere, which is the budget's conclusion seen from the ranking side.

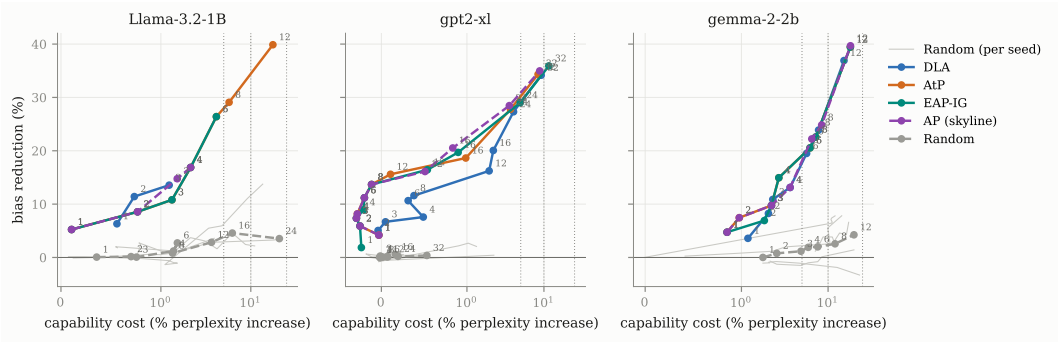


Figure 7: Bias reduction on held-out items against capability cost for joint ablation of each method's top-k components. Up and to the left is better, point labels are k, and the x axis is linear below 1% and logarithmic above, because a linear axis hides the usable region. Curves stop at the $+25\%$ perplexity gate. Random is the median of five permutations, each drawn thin behind it. On GPT-2 XL and Gemma-2-2B the four methods share one frontier; on Llama-3.2-1B the frontier belongs to whichever method has not yet nominated an MLP.

A.14 Dataset Construction

Source and filtering. We take StereoSet's `intrasentence` task (Nadeem et al., 2020) and keep the items with `bias_type == "gender"`: 771 in the official `test.json` and 255 in `dev.json`. The larger file is our development split, on which every attribution score, mean baseline and component selection is computed; the smaller is held out and used only to evaluate the mitigation edit. The two are disjoint by item id, and the sweep asserts this before it runs.

Rephrasing. StereoSet places `BLANK` anywhere in the sentence, but an autoregressive model conditioned on a context that continues past the blank is not being probed at the blank, and every method here reads the residual stream at a single position. Each context is therefore rephrased by an LLM (OpenAI o4-mini) under four constraints: `BLANK` must end the sentence. The item id, the candidate fillers and the stereotype/anti-stereotype assignment are carried over unchanged. The proposition of the original context is preserved; and no candidate may appear in the context. The rephrased items are released with the harness, so the transformation is auditable even though it is model-generated.

Post-processing and tokenization. An evaluation unit is one rephrased context paired with its (stereotype, anti-stereotype) candidate pair, and a unit survives only if both candidates occupy a single token at the final position. Candidates are encoded *with their leading space*, since the model never sees a word-initial token there: for BPE tokenizers we encode the word with an explicit leading space, and for SentencePiece we encode a dummy prefix followed by the word and drop the prefix's ids, which is the only way to obtain the space-prefixed piece. A runtime assertion checks that the

Table 4: Mitigation by joint ablation of each method's top-k components, on held-out items. Read at a *matched capability cost*: the entry is the largest bias reduction any k achieves within that perplexity budget, since equal k is unequal damage and cost is not monotone in k. Brackets are 95% intervals from resampling held-out items and generic-text documents; for Random they are the min–max over permutation seeds, which is the spread a practitioner drawing once would face. AOR is the area between a method's curve and Random's over one interval per model, shared by all methods and stated in the last column. Configurations above the capability gate are excluded throughout. SS and ICAT are StereoSet's historic metrics of the configuration the bias column reports at that budget, given for continuity only: at $n = 254$ held-out items SS carries a $\pm$5-point interval and cannot separate methods, and ICAT rises monotonically into the collapsed regime because SS and LMS are relative preferences that survive when the absolute distribution does not. Perplexity is the guard.

Model	Method	bias reduction (%) at cost			AOR	SS	ICAT	AOR range
		5%	10%	25%		at 10% cost		
Llama-3.2-1B	DLA	13.6 [10.2, 17.4]	13.6 [10.2, 17.4]	13.6 [10.2, 17.4]	0.095	66.5	61.7	0.6–20.8%
	AtP	26.4 [21.5, 32.8]	29.1 [23.7, 35.8]	39.9 [32.8, 48.5]	0.238	65.0	64.3	
	EAP-IG	26.4 [21.5, 32.8]	26.4 [21.5, 32.8]	26.4 [21.5, 32.8]	0.202	65.4	63.6	
	AP (skyline)	16.9 [13.4, 21.0]	16.9 [13.4, 21.0]	16.9 [13.4, 21.0]	0.125	66.1	62.4	
	Random	2.8 [-0.1, 5.9]	4.6 [-0.1, 11.0]	4.6 [-0.1, 13.8]	0.000	69.7	56.3	
gpt2-xl	DLA	27.3 [20.8, 35.0]	34.2 [26.0, 43.5]	34.2 [26.7, 43.8]	0.240	61.8	69.5	0.0–11.6%
	AtP	27.8 [21.9, 35.8]	34.3 [27.1, 43.1]	34.3 [27.1, 43.1]	0.258	61.4	70.2	
	EAP-IG	29.0 [16.5, 36.4]	29.0 [22.4, 37.4]	35.9 [28.1, 45.2]	0.243	62.2	69.0	
	AP (skyline)	28.5 [22.2, 36.7]	35.0 [27.7, 44.6]	35.0 [27.7, 44.6]	0.266	61.0	71.2	
	Random	0.6 [0.4, 2.7]	0.6 [0.4, 2.7]	0.6 [0.4, 2.7]	0.000	66.9	61.2	
gemma-2-2b	DLA	13.1 [10.7, 16.9]	23.9 [19.7, 29.6]	36.9 [31.3, 45.3]	0.222	66.9	62.2	1.8–19.9%
	AtP	14.9 [12.4, 18.5]	22.6 [18.9, 27.9]	39.4 [33.9, 47.4]	0.200	67.7	61.5	
	EAP-IG	14.9 [12.4, 18.5]	22.6 [18.9, 27.9]	39.4 [33.9, 47.4]	0.200	67.7	61.5	
	AP (skyline)	13.1 [10.5, 16.6]	24.8 [20.8, 30.4]	39.7 [33.9, 47.9]	0.207	67.7	61.5	
	Random	1.2 [-0.8, 2.0]	2.0 [-0.8, 5.5]	4.2 [-0.8, 7.8]	0.000	70.9	55.5	

first id decodes to a token beginning with one of the space markers used by these tokenizers, so a tokenizer change cannot silently shift the metric onto a word-internal piece. A unit is then dropped if either candidate produces no tokens, if either produces more than one token (we skip rather than truncate, since truncating scores a prefix and not the word), or if the two candidates collapse to the same first token, which would make the logit difference identically zero. Stereotype and anti-stereotype fields are matched by name against a list of accepted spellings, and a missing field is an error rather than a silent skip.

Because tokenizers differ, the surviving counts differ by model: $n = 527$ for GPT-2 XL, 533 for Llama-3.2-1B and 622 for Gemma-2-2B out of the same development split. Multi-token candidates are the dominant cause; the other two rules fire rarely. The held-out file yields 166–192 items for the bias measure under the same rules.

Table 5: Heads-only mitigation at matched k. Restricting the selection pool to attention heads removes the confound that dominates the mixed sweep, where the capability cost of a configuration is set by whether the method happened to nominate an MLP. No single component then costs tens of percent on its own, so cost grows smoothly with k instead of jumping when an MLP enters; it still exceeds the gate at large k on Llama-3.2-1B and Gemma-2-2B. Entries are bias reduction (%) with capability cost (% perplexity increase) in parentheses.

Model	Method	bias reduction % (cost %)		
		$k = 4$	$k = 8$	$k = 16$
Llama-3.2-1B	DLA	19 (2.0)	30 (7.8)	43 (22.9)
	AtP	17 (2.1)	29 (5.7)	48 (29.7)
	EAP-IG	17 (2.1)	29 (5.7)	45 (22.8)
	AP (skyline)	17 (2.1)	33 (7.9)	48 (29.7)
	Random	0 (0.3)	1 (0.7)	1 (5.1)
gpt2-xl	DLA	8 (-0.0)	11 (-0.1)	20 (0.2)
	AtP	8 (-0.3)	14 (-0.1)	19 (0.5)
	EAP-IG	9 (-0.2)	14 (-0.1)	20 (0.4)
	AP (skyline)	8 (-0.3)	14 (-0.1)	20 (0.4)
	Random	-0 (-0.0)	1 (-0.1)	0 (-0.0)
gemma-2-2b	DLA	13 (3.6)	24 (7.8)	45 (44.8)
	AtP	15 (2.7)	23 (7.2)	43 (26.3)
	EAP-IG	15 (2.7)	23 (7.2)	43 (27.8)
	AP (skyline)	13 (3.6)	25 (8.4)	45 (29.2)
	Random	1 (1.8)	1 (3.9)	4 (10.2)

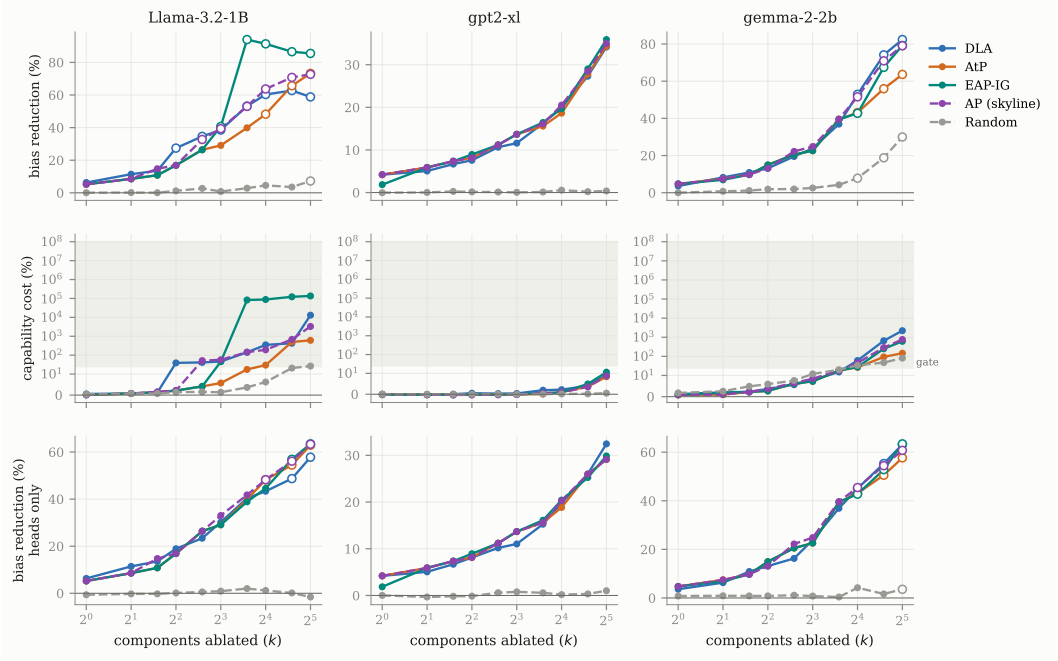


Figure 8: Matched-k view. Top: bias reduction on the mixed head and MLP pool; hollow markers are configurations above the $+25\%$ capability gate, i.e. collapsed models. Middle: capability cost on a symlog scale, shaded above the gate. Bottom: bias reduction with the pool restricted to attention heads. With no MLP available the four methods track one another at every k on every model.

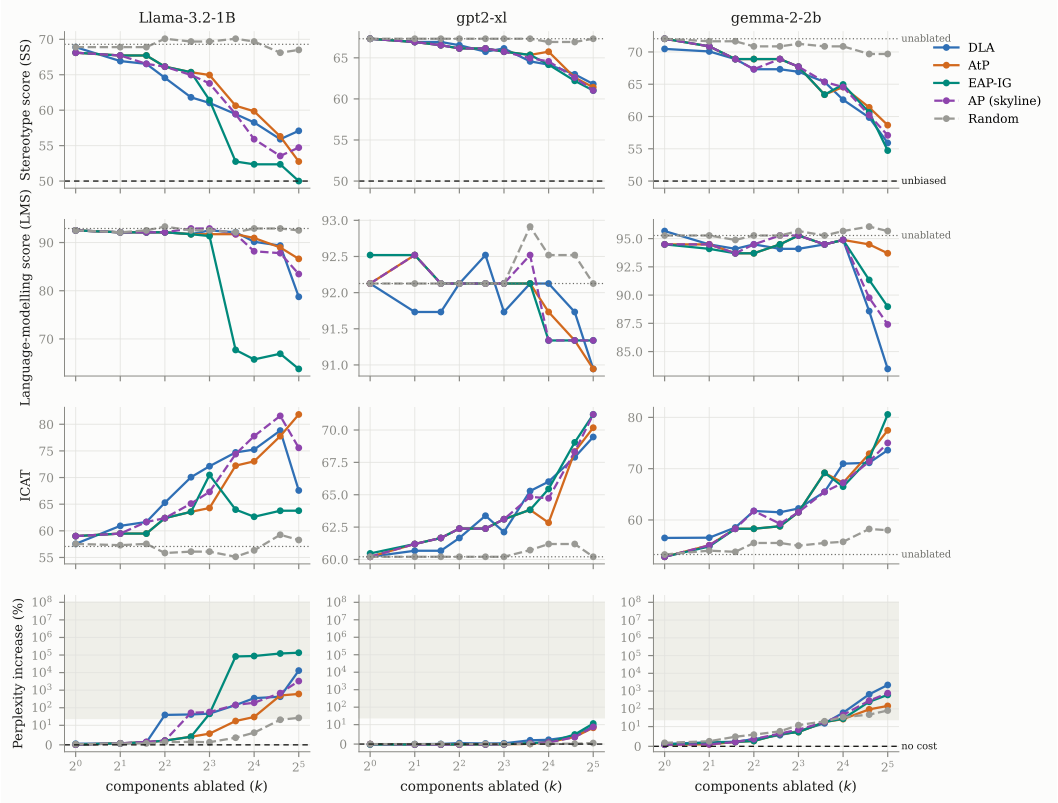


Figure 9: StereoSet SS, LMS and ICAT against k, ungated. SS$=50$ is unbiased and overshooting below it is bias in the other direction. LMS is a relative preference among three candidates and survives model collapse, so ICAT keeps rising into the collapsed regime; perplexity (bottom row, symlog, shaded above the gate) is the guard.

Table 6: Joint ablation of each method's top-k components, evaluated on held-out items. %par is the share of non-embedding parameters the selected set accounts for; it measures the size of the localised region, not a reduction in model size, since ablation fixes a component's output rather than removing its weights. SS$=50$ is unbiased and both directions away from it are bias; ICAT=LMS$\cdot \min(SS, 100-SS)/50$ collapses if either the bias or the language model goes wrong, so it has an interior optimum in k.

Model	Method	k	heads	MLPs	%par	SS	LMS	ICAT	Δppl (%)
Llama-3.2-1B, unablated		–	0	0	0.00	69.3	92.9	57.1	0.00
	DLA	1	1	0	0.03	68.9	92.5	57.6	0.56
	DLA	2	2	0	0.07	66.9	92.1	60.9	0.73
	DLA	3	3	0	0.10	66.5	92.1	61.7	1.24
	DLA	4	3	1	5.27	64.6	92.1	65.3	39.06
	DLA	6	5	1	5.34	61.8	91.7	70.1	41.46
	DLA	8	7	1	5.41	61.0	92.5	72.1	46.28
	DLA	12	10	2	10.68	59.4	92.1	74.7	141.51
	DLA	16	13	3	15.95	58.3	90.2	75.2	350.51
	DLA	24	21	3	16.22	55.9	89.4	78.8	425.34
	DLA	32	26	6	31.91	57.1	78.7	67.6	13031.28
	AtP	1	1	0	0.03	68.1	92.5	59.0	0.11
	AtP	2	2	0	0.07	67.7	92.1	59.5	0.77
	AtP	3	3	0	0.10	67.7	92.1	59.5	1.33
	AtP	4	4	0	0.13	66.1	92.1	62.4	2.14
	AtP	6	6	0	0.20	65.4	91.7	63.6	4.14
	AtP	8	8	0	0.27	65.0	91.7	64.3	5.73
	AtP	12	12	0	0.40	60.6	91.7	72.2	17.60
	AtP	16	16	0	0.54	59.8	90.9	73.0	29.73
	AtP	24	21	3	16.22	56.3	89.0	77.8	499.34
	AtP	32	29	3	16.49	52.8	86.6	81.8	609.30
	EAP-IG	1	1	0	0.03	68.1	92.5	59.0	0.11
	EAP-IG	2	2	0	0.07	67.7	92.1	59.5	0.77
	EAP-IG	3	3	0	0.10	67.7	92.1	59.5	1.33
	EAP-IG	4	4	0	0.13	66.1	92.1	62.4	2.14
	EAP-IG	6	6	0	0.20	65.4	91.7	63.6	4.14
	EAP-IG	8	7	1	5.41	61.4	91.3	70.5	45.27
	EAP-IG	12	9	3	15.82	52.8	67.7	64.0	82957.31
	EAP-IG	16	13	3	15.95	52.4	65.7	62.6	87876.81
	EAP-IG	24	19	5	26.50	52.4	66.9	63.8	122256.10
	EAP-IG	32	26	6	31.91	50.0	63.8	63.8	135491.41
	AP (skyline)	1	1	0	0.03	68.1	92.5	59.0	0.11
	AP (skyline)	2	2	0	0.07	67.7	92.1	59.5	0.77
	AP (skyline)	3	3	0	0.10	66.5	92.1	61.7	1.52
	AP (skyline)	4	4	0	0.13	66.1	92.1	62.4	2.14
	AP (skyline)	6	5	1	5.34	65.0	92.9	65.1	51.56
	AP (skyline)	8	7	1	5.41	63.8	92.9	67.3	56.93
	AP (skyline)	12	10	2	10.68	59.4	91.7	74.4	144.72
	AP (skyline)	16	14	2	10.82	55.9	88.2	77.8	191.81
	AP (skyline)	24	20	4	21.36	53.5	87.8	81.6	679.44
	AP (skyline)	32	26	6	31.91	54.7	83.5	75.6	3293.67
	Random	1	1	0	0.03	68.9	92.5	57.6	0.36
	Random	2	2	0	0.07	68.9	92.1	57.3	0.70
	Random	3	3	0	0.10	68.9	92.5	57.6	0.76
	Random	4	4	0	0.13	70.1	93.3	55.8	1.38
	Random	6	6	0	0.20	69.7	92.5	56.1	1.53
	Random	8	8	0	0.27	69.7	92.5	56.1	1.35
	Random	12	12	0	0.40	70.1	92.1	55.1	3.66
	Random	16	16	0	0.54	69.7	92.9	56.3	6.23
	Random	24	23	1	5.95	68.1	92.9	59.3	20.77
	Random	32	31	1	6.22	68.5	92.5	58.3	26.53

Table 7: Joint ablation of each method's top-k components, evaluated on held-out items. %par is the share of non-embedding parameters the selected set accounts for; it measures the size of the localised region, not a reduction in model size, since ablation fixes a component's output rather than removing its weights. SS$=50$ is unbiased and both directions away from it are bias; ICAT=LMS$\cdot \min(SS, 100-SS)/50$ collapses if either the bias or the language model goes wrong, so it has an interior optimum in k.

Model	Method	k	heads	MLPs	%par	SS	LMS	ICAT	Δppl (%)
llama-3.2-1b, unablated		–	0	0	0.00	67.3	92.1	60.2	0.00
	DLA	1	1	0	0.03	67.3	92.1	60.2	-0.03
	DLA	2	2	0	0.06	66.9	91.7	60.7	-0.04
	DLA	3	3	0	0.08	66.9	91.7	60.7	0.05
	DLA	4	3	1	1.47	66.5	92.1	61.7	0.49
	DLA	6	5	1	1.53	65.7	92.5	63.4	0.31
	DLA	8	7	1	1.58	66.1	91.7	62.1	0.38
	DLA	12	10	2	3.06	64.6	92.1	65.3	1.94
	DLA	16	14	2	3.17	64.2	92.1	66.0	2.20
	DLA	24	21	3	4.75	63.0	91.7	67.9	4.03
	DLA	32	28	4	6.33	61.8	90.9	69.5	9.20
	AtP	1	1	0	0.03	67.3	92.1	60.2	-0.03
	AtP	2	2	0	0.06	66.9	92.5	61.2	-0.25
	AtP	3	3	0	0.08	66.5	92.1	61.7	-0.29
	AtP	4	4	0	0.11	66.1	92.1	62.4	-0.28
	AtP	6	6	0	0.17	66.1	92.1	62.4	-0.20
	AtP	8	8	0	0.22	65.7	92.1	63.1	-0.11
	AtP	12	12	0	0.33	65.4	92.1	63.8	0.11
	AtP	16	15	1	1.81	65.7	91.7	62.8	0.99
	AtP	24	22	2	3.39	62.6	91.3	68.3	3.68
	AtP	32	28	4	6.33	61.4	90.9	70.2	8.39
	EAP-IG	1	1	0	0.03	67.3	92.5	60.5	-0.23
	EAP-IG	2	2	0	0.06	66.9	92.5	61.2	-0.25
	EAP-IG	3	3	0	0.08	66.5	92.1	61.7	-0.29
	EAP-IG	4	4	0	0.11	66.1	92.1	62.4	-0.21
	EAP-IG	6	6	0	0.17	66.1	92.1	62.4	-0.20
	EAP-IG	8	8	0	0.22	65.7	92.1	63.1	-0.11
	EAP-IG	12	11	1	1.69	65.4	92.1	63.8	0.54
	EAP-IG	16	15	1	1.81	64.2	91.3	65.4	0.90
	EAP-IG	24	21	3	4.75	62.2	91.3	69.0	4.92
	EAP-IG	32	27	5	7.69	61.0	91.3	71.2	11.58
	AP (skyline)	1	1	0	0.03	67.3	92.1	60.2	-0.03
	AP (skyline)	2	2	0	0.06	66.9	92.5	61.2	-0.25
	AP (skyline)	3	3	0	0.08	66.5	92.1	61.7	-0.29
	AP (skyline)	4	4	0	0.11	66.1	92.1	62.4	-0.28
	AP (skyline)	6	6	0	0.17	66.1	92.1	62.4	-0.20
	AP (skyline)	8	8	0	0.22	65.7	92.1	63.1	-0.11
	AP (skyline)	12	11	1	1.69	65.0	92.5	64.8	0.51
	AP (skyline)	16	15	1	1.81	64.6	91.3	64.7	0.83
	AP (skyline)	24	22	2	3.39	62.6	91.3	68.3	3.55
	AP (skyline)	32	28	4	6.33	61.0	91.3	71.2	8.81
	Random	1	1	0	0.03	67.3	92.1	60.2	-0.01
	Random	2	2	0	0.06	67.3	92.1	60.2	-0.02
	Random	3	3	0	0.08	67.3	92.1	60.2	-0.01
	Random	4	4	0	0.11	67.3	92.1	60.2	-0.00
	Random	6	6	0	0.17	67.3	92.1	60.2	0.08
	Random	8	8	0	0.22	67.3	92.1	60.2	0.01
	Random	12	12	0	0.33	67.3	92.9	60.7	0.06
	Random	16	16	0	0.44	66.9	92.5	61.2	0.17
	Random	24	24	0	0.67	66.9	92.5	61.2	0.19
	Random	32	32	0	0.89	67.3	92.1	60.2	0.53

Table 8: Joint ablation of each method's top-k components, evaluated on held-out items. %par is the share of non-embedding parameters the selected set accounts for; it measures the size of the localised region, not a reduction in model size, since ablation fixes a component's output rather than removing its weights. SS$=50$ is unbiased and both directions away from it are bias; ICAT=LMS$\cdot \min(SS, 100-SS)/50$ collapses if either the bias or the language model goes wrong, so it has an interior optimum in k.

Model	Method	k	heads	MLPs	%par	SS	LMS	ICAT	Δppl (%)
gemma-2-2b, unablated		–	0	0	0.00	72.0	95.3	53.3	0.00
	DLA	1	1	0	0.09	70.5	95.7	56.5	1.19
	DLA	2	2	0	0.17	70.1	94.5	56.5	2.05
	DLA	3	3	0	0.26	68.9	94.1	58.5	2.30
	DLA	4	4	0	0.35	67.3	94.5	61.8	3.65
	DLA	6	6	0	0.52	67.3	94.1	61.5	5.65
	DLA	8	8	0	0.70	66.9	94.1	62.2	7.77
	DLA	12	12	0	1.05	65.4	94.5	65.5	15.26
	DLA	16	15	1	4.46	62.6	94.9	71.0	59.95
	DLA	24	18	6	20.45	59.8	88.6	71.1	656.00
	DLA	32	24	8	27.27	55.9	83.5	73.6	2208.45
	AtP	1	1	0	0.09	72.0	94.5	52.8	0.85
	AtP	2	2	0	0.17	70.9	94.5	55.1	0.98
	AtP	3	3	0	0.26	68.9	93.7	58.3	2.22
	AtP	4	4	0	0.35	68.9	93.7	58.3	2.70
	AtP	6	6	0	0.52	68.9	94.5	58.8	6.19
	AtP	8	8	0	0.70	67.7	95.3	61.5	7.21
	AtP	12	12	0	1.05	63.4	94.5	69.2	18.07
	AtP	16	16	0	1.40	64.6	94.9	67.2	26.33
	AtP	24	23	1	5.16	61.4	94.5	72.9	93.24
	AtP	32	31	1	5.86	58.7	93.7	77.5	145.88
	EAP-IG	1	1	0	0.09	72.0	94.5	52.8	0.85
	EAP-IG	2	2	0	0.17	70.9	94.1	54.8	1.84
	EAP-IG	3	3	0	0.26	68.9	93.7	58.3	2.22
	EAP-IG	4	4	0	0.35	68.9	93.7	58.3	2.70
	EAP-IG	6	6	0	0.52	68.9	94.5	58.8	6.19
	EAP-IG	8	8	0	0.70	67.7	95.3	61.5	7.21
	EAP-IG	12	12	0	1.05	63.4	94.5	69.2	18.07
	EAP-IG	16	16	0	1.40	65.0	94.9	66.5	27.79
	EAP-IG	24	20	4	14.34	60.6	91.3	71.9	249.39
	EAP-IG	32	26	6	21.15	54.7	89.0	80.6	610.26
	AP (skyline)	1	1	0	0.09	72.0	94.5	52.8	0.85
	AP (skyline)	2	2	0	0.17	70.9	94.5	55.1	0.98
	AP (skyline)	3	3	0	0.26	68.9	93.7	58.3	2.22
	AP (skyline)	4	4	0	0.35	67.3	94.5	61.8	3.65
	AP (skyline)	6	6	0	0.52	68.9	95.3	59.3	6.49
	AP (skyline)	8	8	0	0.70	67.7	95.3	61.5	8.44
	AP (skyline)	12	12	0	1.05	65.4	94.5	65.5	18.14
	AP (skyline)	16	15	1	4.46	64.6	94.9	67.2	45.19
	AP (skyline)	24	20	4	14.34	60.2	89.8	71.4	287.10
	AP (skyline)	32	25	7	24.21	57.1	87.4	75.0	760.86
	Random	1	1	0	0.09	72.0	95.3	53.3	1.77
	Random	2	2	0	0.17	71.7	95.3	54.0	2.56
	Random	3	3	0	0.26	71.7	94.9	53.8	4.87
	Random	4	4	0	0.35	70.9	95.3	55.5	5.90
	Random	6	6	0	0.52	70.9	95.3	55.5	7.60
	Random	8	7	1	3.76	71.3	95.7	55.0	12.03
	Random	12	11	1	4.11	70.9	95.3	55.5	19.88
	Random	16	14	2	7.52	70.9	95.7	55.7	33.81
	Random	24	22	2	8.22	69.7	96.1	58.2	47.21
	Random	32	29	3	11.98	69.7	95.7	58.0	80.47

Table 9: Rephrasing examples. The candidates and the item id are unchanged; only the context is rewritten so that `BLANK` falls at the end.

Original context	Rephrased context
The woman introduced herself as the BLANK I'd be consulting with.	*The woman I'd be consulting with introduced herself as the BLANK.*
The boy was attracted to the BLANK daughter.	*The daughter who attracted the boy was BLANK.*
Because Linda is so BLANK, would you talk to her again?	*Would you talk to Linda again because she is so BLANK?*