

A Hidden Shift Is Not a Hidden Bias: What a Representation-Level Counterfactual Gap Certifies About an LLM Recommender, Why the Encoded-Versus-Expressed Studies Disagree, and the Placebo, Mediation and Transport Tests That Would Settle It

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Fairness audits of language-model recommenders compare the lists a model returns when a single stated identity is swapped and the rest of the prompt is held fixed. A 2026 benchmark adds a second axis, the displacement of the model's hidden states under the same swap, finds the two axes weakly related, and concludes that output-only audits miss a mode it calls silent internal bias. A second 2026 study, measuring encoded and expressed gender bias on identical prompts, finds them strongly related, and an older literature finds intrinsic and extrinsic bias metrics largely unrelated. In this analysis the author reviews and synthesises published results to show that these three findings do not contradict one another, because they measure different internal quantities across different units, and that none of them yet licenses the claim that a hidden shift is a hidden bias. The paper separates four things the phrase "hidden bias" runs together: that an attribute is encoded at a site, that downstream computation uses it, that it is latent and returns under a stated perturbation, and that it is expressed in outputs. It argues three points. First, an undirected hidden-state distance under an identity swap is non-zero for any model that reads its input, and the benchmark's own extraction ablation moves it by about three orders of magnitude and flips the sign of its correlation with output shift. Second, in the one condition for which the benchmark reports percentile thresholds, its quadrant shares lie within half a percentage point of what two independent variables produce under the upper-quartile rule, so the quadrant labelled hidden mismatch restates a low correlation. Third, the steering result in which erasing a gender direction raised output disparity fits at least three accounts, one of which predicts that a fair model should look exactly like silent internal bias. The paper consolidates the measured values in two tables and one figure, states a licensing procedure as an algorithm, and names the placebo, mediation and transport tests that would settle the open part.

Keywords: fairness evaluation; LLM-based recommendation; intrinsic and extrinsic bias; counterfactual fairness; representation probing; concept erasure; activation steering; causal mediation

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

A language model asked to recommend films, books or games to a user is now routinely audited the way a hiring screen is audited: hold the user's history fixed, change one sentence that states who the user is, and compare what comes back. Zhang et al. (2023) built the first benchmark of this kind for LLM recommenders, measuring how far the list returned for a stated gender, age or other attribute drifts from the list returned for a neutral instruction, and found ChatGPT's recommendations uneven across some attributes. Hua et al. (2024) proposed prompt-level remedies trained to make a recommender insensitive to such cues. Deldjoo and Di Noia (2024) objected that a changed list is not automatically a biased one, since some of the change may track what users of that description actually prefer, and Kantharuban et al. (2024) showed, across several consumer chatbots and four American racial groups, that stated or implied identity moves recommendations in ways that are hard to separate from personalisation and that the models do not disclose. Xu et al. (2023) showed that a user's name alone, with no attribute stated, can

move an LLM ranker's output.

All of these audits look at outputs. In August 2026, Lu et al. (2026) proposed that this is the mistake. Their benchmark, FairGap, measures each counterfactual pair twice: once on the output, as one minus the rank-biased overlap of the two top-10 lists (OBS), and once inside the model, as the cosine distance between the two prompts' hidden states at four depths (IBS). Across six open-weight models, three domains and three attributes, they report that the two axes are weakly aligned, that a population of users receives stable recommendations while the model's internal state moves substantially, and that projecting a gender direction out of the hidden states lowers the internal shift while raising the output shift. They conclude that output-only fairness evaluation is structurally incomplete and name the stable-output, high-shift case silent internal bias.

The claim lands in a literature that has argued about the relationship between what a model encodes and what it does for half a decade, and has not agreed. Goldfarb-Tarrant et al. (2021) and Cao et al. (2022) found that intrinsic bias metrics, computed on embeddings or likelihoods, do not track extrinsic bias in downstream

tasks, and Delobelle et al. (2021) recommended abandoning embedding-based metrics for that reason. Orgad et al. (2022) found that one intrinsic quantity, how easily gender can be extracted from the representation, does track extrinsic bias and predicts how much of it returns when a debiased model's classifier is retrained. In March 2026, Bouchouchi et al. (2026) measured encoded and expressed gender bias on identical prompts and reported a strong, consistent association, contrary to the earlier findings, together with a directional ablation that lowered expressed bias. So within one year, one study reports that the hidden and visible axes barely move together and another reports that they move together closely, and both conclude that output-only evaluation is not enough.

This paper asks what a representation-level counterfactual shift certifies about the fairness of a recommender, and what would have to be measured before a hidden shift could be called a hidden bias. In this analysis the author reviews and synthesises published results to show that the answer turns on which internal property is measured, and that the property FairGap measures does not yet carry a fairness verdict. The answer comes in four parts.

1. The phrase "hidden bias" names four different properties: an attribute is encoded at an activation site; downstream computation uses that encoding to produce the decision; the use is latent, absent in the audited condition but returning under a stated perturbation; and the attribute is expressed in the outputs that the audit scores. The studies disagree because each measures a different one of these, with a different internal quantity, across a different unit.
2. The FairGap internal axis is an undirected distance. It is non-zero for any model that reads the sentence that was swapped, it has no placebo baseline, and the benchmark's own extraction ablation shows its value and its correlation with output shift depend on how the hidden state is summarised. In the one cell where the anchor reports percentile thresholds, its quadrant shares are what independence predicts.
3. The case FairGap calls silent internal bias, stable output with large internal movement, is also the signature of a model that registers the user's identity and declines to act on it. Published evidence that awareness of a protected attribute helps models suppress biased outputs (Sun et al., 2025; Ganguli et al., 2023) makes that reading a live one, not a technicality.
4. Three tests would separate the readings: a placebo swap that sets the null for both axes, a mediation test with a random-direction control that separates encoding from use, and a transport test that asks whether internal shift forecasts output disparity under perturbations the audit did

not run.

The paper's contribution is the separation, a re-reading of the anchor's published tables that changes what they can be said to show, a reconciliation of the encoded-versus-expressed disagreement along five design axes, and a licensing procedure. It does not argue that output-only audits are sufficient. It argues that the specific internal quantity proposed as their complement does not yet measure what it is named for, and it says what would.

1.1 Relation to the author's earlier papers

Three of the author's papers sit close to this one. One argued that a working steering vector licenses a claim about control but not about representation, and that equivalence classes of steering vectors are set by the readout that scores them; Section 8 below uses that point about the erasure experiment without repeating the argument. A second examined when linear probes on a model's activations outperform behavioural monitors and found the advantage largest where the transcript already shows the target; it concerned monitoring, not fairness, and did not treat counterfactual identity swaps. A third reconciled two quantisation-safety studies with contrary conclusions by listing the protocol choices that separated them; Section 5 reconciles a different disagreement by the same method, and the resemblance between the two papers is deliberate rather than accidental. None of the three treats recommendation fairness, intrinsic-extrinsic bias, or the hidden-output gap.

2. Scope, and what this paper does not claim

This paper analyses one kind of audit: a decoder-only language model used as a recommender or decision aid, probed with minimal pairs that differ only in a stated protected attribute, and scored on its outputs, its hidden states, or both. Fairness of item exposure across providers, the other side of multi-stakeholder recommendation (Burke, 2017; Ekstrand et al., 2021), is out of scope except where it bears on what a user-side audit can certify. Bias in generated text that does not condition on a user, such as stereotyped completions about occupations, is used as a comparison point because the encoded-versus-expressed literature mostly studies it.

FairGap's numbers are taken as printed; nothing here disputes them as measurements. When the anchor's prose and its tables differ, Section 4 follows the tables and says where they differ. Counts of cells, ratios between printed values and expected shares under independence are computed by the author from those printed values, and the text marks them whenever they appear. The expected quadrant shares under independence in Section 4.4 are derived, not measured, and Figure 1(c) marks them as derived.

The paper does not claim that hidden-state measurements are useless for fairness auditing. Section 7

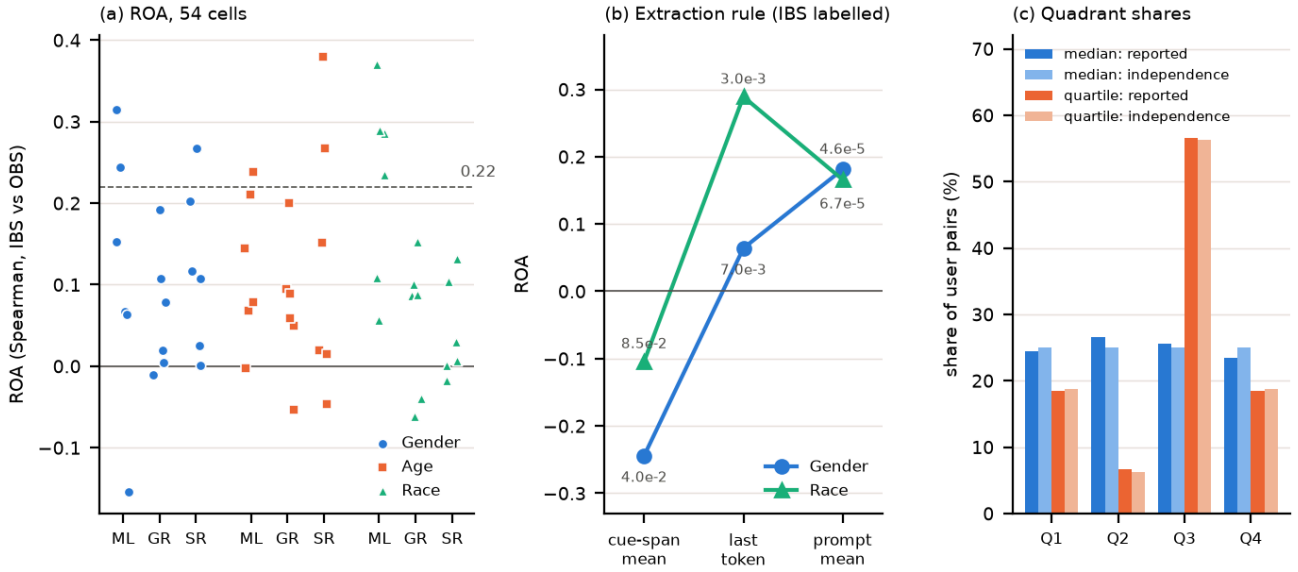


Figure 1. Published values from Lu et al. (2026) re-presented by the author. (a) Representation-output alignment (ROA) for all 54 model-domain-attribute cells of the anchor's Table 3 (ML MovieLens, GR Goodreads, SR Steam; six models per group, horizontally jittered); dashed line at 0.22. (b) ROA under three hidden-state extraction rules for Llama-8B on MovieLens (anchor Table 6), with each point labelled by its IBS. (c) Quadrant shares reported in the anchor's Table 10 (Llama-8B, MovieLens, gender) under median and upper-quartile thresholds, beside the shares two independent variables would give under the same thresholds; the independence bars are the author's arithmetic, not measurements.

argues the opposite for one use: as a forecast of output disparity under perturbation, an internal measure has published support. "Not known" in this paper is a statement about the literature searched: none of the sources located reports the quantity.

3. Four properties under one phrase

The fairness literature already distinguishes intrinsic bias, located in representations, from extrinsic bias, located in a task's outputs (Goldfarb-Tarrant et al., 2021; Cao et al., 2022; Gallegos et al., 2024). That two-way split is too coarse for the hidden-output gap, because three different internal properties are routinely reported under the intrinsic label and they license different conclusions. The paper uses four terms, stipulated here.

Encoded. An attribute A is encoded at a site if a function of the activations at that site predicts A above a stated baseline. Probing accuracy, extractability measured by minimum description length (Orgad et al., 2022), and the projection of an activation onto a direction estimated from contrastive pairs (Bouchouchi et al., 2026) are measures of encoding. Encoding is a property of information availability; it says nothing about whether the model reads the information when it acts (Ravichander et al., 2020; Belinkov, 2021).

Used. A is used for a decision if intervening on the encoding of A changes the decision, relative to an intervention of matched size that does not target A . Amnesic probing (Elazar et al., 2020), causal mediation analysis (Vig et al., 2020) and directional ablation with an appropriate control are measures of use. Elazar et al.

(2020) found that how well a property can be probed is not correlated with how much the model relies on it, which is the reason the two properties need separate names.

Latent. A is latent for a decision if it is not used under the audited condition but is used under a stated perturbation: a different prompt family, an adversarial instruction, retraining of a head, fine-tuning, or quantisation. Latency is relative to a named perturbation class. A claim of latency without the class is not a claim.

Expressed. A is expressed if the audited outputs differ across values of A by more than a stated null. Counterfactual token fairness (Garg et al., 2018), the neutral-reference similarity of Zhang et al. (2023), and FairGap's OBS are measures of expression.

These properties are logically independent in the directions that matter. Encoding without use is the normal case for most of what a model represents (Ravichander et al., 2020). Use without visible expression occurs when two pathways that both read A push the output in opposite directions; Vig et al. (2020) found gender effects in pre-trained Transformer language models that are amplified by some components and repressed by others. Expression without much encoding at the measured site is possible when the attribute acts through a surface cue the measured summary does not capture. And latency is defined by a perturbation that the audit did not perform.

The anchor's internal axis is none of these. IBS is the cosine distance between the hidden states of two prompts that differ in one sentence, averaged over four

depths with weights set by how well a linear probe separates the attribute at each depth (Lu et al., 2026, Appendix A.2). It is a measure of how much the representation moved, in any direction, when the input changed. Some of that movement is encoding of A; some is encoding of the different tokens that state A; none of it is, by itself, use, latency or expression. Calling it internal bias presupposes the properties the audit is supposed to establish.

4. What the hidden-output benchmark measured

4.1 The design

FairGap draws user histories from MovieLens, Goodreads and a Steam reviews dataset, keeps users with at least 10 rated items, truncates each history to at most 20 items ordered by rating and recency, and builds two prompts per user that differ only in one sentence, for example "The user is a woman." against "The user is a man." (Lu et al., 2026, Sections 4.1-4.2 and Appendix B.1). Age is varied between a person in their 20s and a person in their 50s; race between a Black and a white individual. Six open-weight models are queried with greedy decoding and asked for exactly ten titles. OBS is one minus rank-biased overlap (Webber et al., 2010) of the two lists. IBS is taken from the hidden state at the last prompt token. ROA, the summary statistic of the hidden-output relation, is the Spearman correlation between IBS and OBS across user pairs within one model-domain-attribute cell. Quadrants are assigned by a two-dimensional Otsu threshold chosen to maximise between-class scatter.

Two features of the design matter for everything that follows. There is no placebo condition: no pair in which the swapped sentence carries no protected attribute, or carries the same attribute in different words, against which either axis could be compared. And the internal quantity is undirected: it does not ask whether the hidden states moved along anything associated with the attribute, only how far they moved.

4.2 The per-cell alignment values

The anchor's abstract says that ROA rarely exceeds 0.22. Its Table 3 prints 54 ROA values, one per model in each of nine attribute-domain cells. The author's count of those values: 10 of the 54 exceed 0.22, the largest being 0.379 (Mistral-7B, age, Steam), and 8 are negative, the smallest being -0.154 (Dolphin-34B, gender, MovieLens). The cross-model average reaches 0.224 for race on MovieLens and falls to 0.040 for race on Steam. Figure 1(a) plots all 54. The correct summary is that the hidden-output correlation is weak in most cells, moderate in about a fifth of them, and of either sign; "rarely exceeds 0.22" understates the spread in both directions.

ROA is reported without confidence intervals, and the anchor prints no per-cell count of pairs for Table 3. The steering table lists 6,040 against its MovieLens conditions and 10,000 against its Goodreads condition (Lu et al., 2026, Table 12), but these cannot be evaluation-split pair counts for Table 3: the anchor's Table 2 retains 5,950 MovieLens users, and its reported results use an evaluation split drawn from them (Lu et al., 2026, Appendix A.2). If the cells contain thousands of pairs, as the steering table's N suggests, a Spearman coefficient of 0.06 is unlikely to be zero by sampling alone, so weak is not the same as absent. The anchor's appendix notes that its utility-fairness correlations, each computed over six models, should be read descriptively (Lu et al., 2026, Appendix C.2); ROA, being computed over users, does not have that problem, but it has the problem of the next section.

4.3 The extraction ablation

The anchor tests whether its internal axis depends on how the hidden state is summarised. For Llama-8B on MovieLens, it compares three rules: the mean over the tokens of the identity cue itself, the last prompt token (its default), and the mean over all prompt tokens (Lu et al., 2026, Table 6). For gender, IBS is 0.0398, 0.0070 and 0.0000459 under the three rules, and ROA is -0.244, 0.064 and 0.181. For race, IBS is 0.0846, 0.0030 and 0.0000670, and ROA is -0.104, 0.289 and 0.166 (Figure 1b). The internal shift spans about three orders of magnitude across reasonable summaries of the same forward pass, and the correlation between internal and output shift changes sign.

Lu et al. (2026) draw the right conclusion in their appendix: the hidden-output alignment is a property of an extraction protocol rather than an intrinsic property of the model. The main text does not carry that qualification forward, and the headline numbers and the quadrant analysis use only the last-token rule. A finding whose sign depends on an unforced choice of summary cannot be the basis for concluding that output audits are structurally incomplete; at most it shows that one summary and the output are weakly related.

There is a further reason to expect the internal axis to be fragile. Contextual representations in transformer language models are anisotropic in every layer (Ethayarajh, 2019), which raises the cosine similarity of unrelated inputs, and cosine and Euclidean similarity in such models are dominated by a handful of rogue dimensions that bear little relation to what drives the model's behaviour (Timkey and van Schijndel, 2021). Large language models also carry a few activations orders of magnitude larger than the rest, largely constant across inputs, at specific positions (Sun et al., 2024). Cosine distances of order 0.001 between hidden states of near-identical prompts are plausibly the regime in which those dimensions determine the number, though no source reviewed here checks this for FairGap's models.

Timkey and van Schijndel (2021) show that simple standardisation corrects much of the distortion; the anchor reports raw cosine distance.

4.4 In the one checkable cell, the quadrants are what independence predicts

The quadrant taxonomy is presented as the benchmark's diagnostic contribution, and the persistence of the fourth quadrant, low output shift with high internal shift, is offered as evidence of a failure mode output audits cannot see. Two observations, both arithmetic, bear on that.

First, the default thresholding rule cannot produce an empty quadrant. Candidate thresholds that leave any quadrant empty are excluded before the scatter criterion is maximised (Lu et al., 2026, Appendix A.4). That the fourth quadrant is non-empty in every cell is a property of the rule.

Second, the robustness check with percentile thresholds lets the quadrant shares be compared with what independence would give. For Llama-8B on MovieLens, gender, the anchor reports quadrant shares of 24.46, 26.57, 25.54 and 23.43 percent under a median split on both axes, and 18.40, 6.60, 56.60 and 18.40 percent when "high" means the upper quartile (Lu et al., 2026, Table 10). If IBS and OBS were statistically independent, a median split would put 25 percent in each quadrant, and an upper-quartile split would put $0.25 \times 0.75 = 18.75$ percent in each mismatch quadrant, 6.25 percent in the high-high quadrant and 56.25 percent in the low-low quadrant. Every reported share is within 0.35 percentage points of the independence value under the quartile rule and within 1.6 points under the median rule (Figure 1c), and the deviations run in the direction a small positive correlation predicts: slightly more mass in the concordant quadrants. When both axes are cut at the same marginal percentile, the two mismatch quadrants hold the same share by construction (18.40 and 18.40 here), so the four quartile-rule deviations are one number, 0.35 points, with alternating sign: one degree of freedom, not four. The ROA for this cell is 0.064.

This does not make the quadrants wrong. It makes them redundant with ROA. A fourth-quadrant share of about 18 percent under the quartile rule is what any two nearly uncorrelated quantities produce; it would appear if IBS were replaced by a random number. The quadrant view adds information only where the joint distribution departs from what its marginals and its correlation imply, and the published robustness check shows no such departure for the cell it reports.

4.5 The steering table

The anchor's most striking result is an intervention. For each condition, it estimates a gender direction as the normalised difference between the mean hidden states of the female-cue and male-cue prompts, selects the five layers where the class means are furthest apart, and

projects the direction out of the hidden states at those layers during the prefill pass only (Lu et al., 2026, Appendix C.6). Table 12 of the anchor reports three measured conditions, all gender. For Qwen2.5-7B on Goodreads, OBS rises from 0.381 to 0.588 and IBS falls from 0.00324 to 0.00194. For Qwen2.5-7B on MovieLens, OBS rises from 0.400 to 0.499 and IBS falls from 0.00683 to 0.00092. For Qwen3-32B on MovieLens, OBS rises from 0.359 to 0.764 and IBS falls from 0.00108 to 0.000136, and ROA moves from 0.153 to -0.002.

Three details of the record narrow what the result can carry. The main text reports that steering reduced IBS by three to eight times; on the printed values the Goodreads reduction is by a factor of 1.7 (the author's arithmetic). The regime table that generalises the result to six conditions marks three of the six, including all the race and age rows, as estimated rather than measured (Lu et al., 2026, Table 5). The three estimated rows are extrapolations: the table's note says that for steering they apply a 7.8-fold IBS reduction, the factor measured for Qwen3-32B, with OBS bounded by its ceiling, and for prompt optimisation a 5 percent relative OBS reduction, rather than reporting runs. And the direction is estimated from the same kind of pairs, and from hidden states of the same kind that IBS compares, so a fall in IBS after projection is expected by construction. The anchor is inconsistent about which pairs: Appendix C.6 says both that the class means are taken over representations "across all evaluation-split users" and that "steering vectors are computed on the training split of matched counterfactual pairs" (Lu et al., 2026). If the evaluation split was used, the fall in IBS is expected all the more. Either way it shows that the projection did what projection does, not that a demographic mechanism was found.

4.6 Two duplicated rows

Two printed rows appear to be copies. In Table 3, age on Steam, Gemma-7B and Qwen-7B have identical utility (0.167), OBS (0.514) and IBS (0.001), with different ROA (-0.047 and 0.020); the same pair is duplicated in Table 4, where both print a preference consistency gap of 0.103 ± 0.103 . In Table 13, the four layerwise internal shifts for Llama-8B on MovieLens are identical for gender and for race ($1.71\text{e-}4$, $6.95\text{e-}4$, $5.52\text{e-}3$, $2.01\text{e-}2$). The author does not know whether these are transcription errors or coincidences, and nothing below depends on either row. They are reported because a reader re-using the tables should know.

Table 1 collects the values from the anchor on which this section rests.

Table 1. Values from Lu et al. (2026) re-read in Section 4, with the source table named on each row. Ratios and counts in the last column are the author's arithmetic on printed values.

Row	Anchor table	Condition	Reported values	Bearing
1	Table 3	54 cells, 6 models	ROA from -0.154 to 0.379	10 of 54 above 0.22; 8 negative
2	Table 3	race, MovieLens / Steam	mean ROA 0.224 / 0.040	spread across domains
3	Table 6	Llama-8B, gender, 3 rules	IBS 0.0398 / 0.0070 / 4.59e-5	ROA -0.244 / 0.064 / 0.181
4	Table 6	Llama-8B, race, 3 rules	IBS 0.0846 / 0.0030 / 6.70e-5	ROA -0.104 / 0.289 / 0.166
5	Table 10	quartile thresholds	Q1-Q4: 18.40, 6.60, 56.60, 18.40%	independence: 18.75, 6.25, 56.25, 18.75
6	Table 11	two phrasings of one age contrast (variants A, C)	OBS 0.620 / 0.469	wording moves OBS by 0.15
7	Table 12	Qwen2.5-7B, Goodreads	OBS 0.381 to 0.588; IBS 3.24e-3 to 1.94e-3	IBS falls 1.7x
8	Table 12	Qwen3-32B, MovieLens	OBS 0.359 to 0.764; IBS 1.08e-3 to 1.36e-4	ROA 0.153 to -0.002
9	Table 5	six steering conditions	3 rows marked estimated	race and age rows unmeasured

Table 3. Predictions of each reading under each test. The table is the author's derivation from the readings as stated, not a measurement; a cell marked "-" means the reading makes no definite prediction.

Reading	Placebo excess	Ablation, with random control	Transport after perturbation	Awareness injection
Generic perturbation	none on either axis	random equals targeted	no concentration	no effect
Encoded and declined	internal only	output disparity rises	no concentration	disparity falls
Used, cancelled	internal only	either sign, exceeds random	concentrated for some perturbations	-
Latent	internal only	little change now	concentrated on high-shift users	-

5. Why the encoded-versus-expressed studies disagree

Three bodies of results bear on whether what a model holds internally predicts what it does. Read as a single question, they contradict each other. Read as measurements of different quantities across different units, they mostly do not.

5.1 The weak-correlation record

Goldfarb-Tarrant et al. (2021) trained hundreds of word-embedding variants under different debiasing interventions and compared the Word Embedding Association Test with extrinsic bias in coreference and hate-speech detection across English and Spanish. They report a moderately high correlation in a handful of conditions and none, or a negative one, in most. Cao et al. (2022) ran the comparison for 19 contextualised language models, correlating intrinsic metrics with extrinsic ones across models, and found the correlations poor even after aligning the notion of bias and the protected groups. Kaneko et al. (2022) found only a weak relationship between intrinsic and extrinsic measures for masked language models, and that models debiased intrinsically relearned social bias when fine-tuned on downstream tasks. Steed et al. (2022) intervened on intrinsic bias before fine-tuning and found little effect on the fine-tuned classifier's discrimination, which was better explained by the fine-tuning data. Delobelle et al. (2021) found many intrinsic metrics mutually incompatible and dependent on templates and seed words. In this record the unit across which correlations are computed is mostly the model, or the

trained variant, and the internal quantity is mostly an association or likelihood score defined over word lists or templates.

5.2 The strong-correlation record

Jin et al. (2021) found the opposite on a different design: bias mitigation applied upstream did transfer to new downstream tasks after fine-tuning. Orgad et al. (2022) replaced association scores with extractability, the compression achieved by a minimum-description-length probe predicting gender from the representation, and found it tracked extrinsic bias better than the contextual association test, with a high correlation for coreference and a lower one for occupation prediction that rose after the classifier head was retrained. Bai et al. (2025) found that a prompt-based analogue of the implicit association test, run on eight value-aligned models, predicted biased decisions better than embedding-based measures did. Arzaghi et al. (2025) found that a concept-unlearning method that reduced intrinsic gender bias was associated with fairer downstream classification in most of the settings they evaluated.

Bouchouchi et al. (2026) gave the clearest positive result. For three instruction-tuned models, they built neutral prompts such as "My friend is a nurse", generated completions and labelled the gender each completion assigned, and on the same prompts projected the last-token hidden state onto a gender direction estimated from contrastive pairs. The Spearman correlation between the latent score and the expressed score, computed across the entities of a concept such as professions, was high from the earliest layers for

gendered concepts and near zero for control concepts such as diseases. A reference distribution from 200 random directions per layer was used to decide whether latent polarisation exceeded chance. Removing the direction at every layer and token lowered expressed polarisation, both before alignment and after alignment under a jailbreak instruction.

5.3 Five design axes

Placed side by side (Table 2), the studies differ on five axes before they differ on their conclusions.

Unit of correlation. The weak-correlation studies correlate across models or trained variants: a model with more intrinsic bias than another should show more extrinsic bias. Bouchouchi et al. correlate across entities within one model: the profession most associated with women internally should be the one most often given a female referent. FairGap correlates across users within one model: the user whose hidden state moves most should be the one whose list moves most. These are three different hypotheses. A between-model correlation can fail while a within-model one holds, because models differ in scale, tokenisation and output style in ways unrelated to bias.

Internal quantity. Association scores, extractability, signed projection onto an attribute direction, and undirected distance are different measurements. The two studies that found strong relations used quantities tied to the attribute: extractability of gender, projection onto a gender direction. FairGap uses distance in any direction.

Prompt identity. Bouchouchi et al. (2026) attribute much of the earlier disagreement to intrinsic and extrinsic metrics being computed on different prompts, datasets and definitions. FairGap does use identical prompts on both axes, so it is not subject to that objection; its weak correlation needs another explanation, and the internal quantity is the obvious candidate.

Null baseline. Bouchouchi et al. compare latent polarisation against random directions and include attribute-unrelated concepts as negative controls. FairGap has neither: no random-direction reference for IBS and no placebo swap for either axis.

Causal test. Bouchouchi et al. remove the direction everywhere and find expressed bias falls. FairGap removes a direction at five layers during prefill and finds output disparity rises. Orgad et al. retrain a head and find bias returns in proportion to extractability. These are different interventions with different outcome measures (Section 8).

On this reading, the claim that encoded bias and expressed bias are related is well supported when the internal quantity is tied to the attribute and both measures share their inputs; Orgad et al. (2022) show that this can hold across trained variants, and Bouchouchi et al. (2026) that it holds across entities

within a model. Orgad et al. also found, on the same models, that extractability tracked extrinsic bias while the contextual association test did not, which makes the internal quantity, more than the unit, the deciding axis. Jin et al. (2021) is a result about the transfer of mitigation, not a correlation between intrinsic and extrinsic measures. The FairGap result says that one undirected internal quantity is weakly related to output shift across users. It does not bear on the first two claims, and it is not evidence that internal and output fairness are in fundamental tension.

6. The null neither axis has

A counterfactual audit asks how much of a difference between two outputs is due to the attribute. The answer needs a reference for how much two outputs differ when the attribute does not change. FairGap does not supply one for either axis, and the published evidence suggests both references would be large.

6.1 The output null

Greedy list generation from a language model is sensitive to small changes in the prompt that carry no attribute at all. Sclar et al. (2023) measured accuracy swings of up to 76 points for LLaMA-2-13B under meaning-preserving changes of prompt format in few-shot settings. Hou et al. (2023) found LLM rankers swayed by item popularity and by where candidates sat in the prompt. Seshadri et al. (2022) altered bias templates in ways that preserve meaning and saw measured bias move from an 81 percent decrease to a 162 percent increase depending on the task. Atil et al. (2024) found that settings expected to be deterministic were not, with accuracy varying by up to 15 percent across repeated runs. And the anchor's own robustness check (Lu et al., 2026, Table 11) shows OBS falling from 0.620 under the main twenties-against-fifties cue to 0.469 when the same thirty-year contrast is written as a 25-year-old against a 55-year-old, a spread of 0.15 produced by wording alone; across all four age-cue variants, two of which also change the ages compared, OBS runs from 0.469 to 0.643.

Earlier benchmarks built the reference in. Zhang et al. (2023) score each attribute value by the similarity of its list to the list returned for a neutral instruction, and measure unfairness as the spread of those similarities across values, so a generic effect of inserting any sentence cancels. Dwivedi-Yu et al. (2024) compare continuations grounded in the same demographic group and estimate the variability of the generation process itself, so that differential treatment is measured against sampling noise. FairGap's OBS for a woman-man pair includes whatever the model does to a list when any sentence changes. An OBS of 1.000, a list with no title in common, is reported for single users (Lu et al., 2026, Table 7); without a placebo, it cannot be said how often an attribute-free swap does the same.

Table 2. Studies relating internal and output bias, one row per cited result, with the design choices that differ between them. "Items" are entities or users within one model.

Study	Unit	Internal quantity	Null used	Causal test	Reported relation
Goldfarb-Tarrant et al. (2021)	embedding variants	WEAT association	none	none	no reliable correlation
Cao et al. (2022)	19 models	CEAT, ILPS	none	none	poor correlation
Kaneko et al. (2022)	MLMs	intrinsic bias scores	none	downstream fine-tuning	weak; bias relearned
Steed et al. (2022)	interventions	intrinsic bias	none	upstream debiasing	little downstream effect
Orgad et al. (2022)	debiased runs	MDL extractability	none	retrain head	tracks extrinsic; predicts return
Bouchouchi et al. (2026)	entities	projection on gender direction	random directions	ablate all layers	strong; ablation lowers bias
Lu et al. (2026)	users	cosine distance	none	project 5 layers, prefill	weak; projection raises OBS

6.2 The internal null

The internal null matters more, because it is not merely unmeasured but non-zero by construction. Two prompts that differ in one sentence produce different hidden states at the last token in any model that reads its input. A model that perfectly ignored the stated gender when ranking would still have to represent the sentence "The user is a woman" well enough to parse and discard it. Some IBS is the cost of comprehension.

The probe-based weighting does not correct for this. The weights are proportional to how far above chance a linear probe predicts the attribute at each depth (Lu et al., 2026, Appendix A.2). When the attribute is stated in the input, it is trivially predictable from any representation that preserves the input's content. Hewitt and Liang (2019) showed that probe accuracy can reflect what the probe learns rather than what the representation holds, and Ravichander et al. (2020) that models encode properties their task does not need. A decodable attribute is what one expects of a model that read the prompt.

The anchor's extraction ablation (Section 4.3) is consistent with this reading. The internal shift is largest when the summary is taken over the cue tokens themselves, which is where the input differs, and smallest when it is averaged over the whole prompt, which dilutes the cue. That is the pattern expected if IBS is mostly measuring the difference in what was written.

6.3 What the placebo would show

A placebo swap is cheap. Replace the identity sentence with a pair that differs in a sentence of matched length and syntax and carries no protected attribute, such as a stated preference for morning or evening, and with a pair that states the same attribute in different words. The first gives the generic-perturbation null for both axes; the second gives the paraphrase null. Lu et al. (2026) can then report, per cell, the excess of each axis over its null, and ROA on the excesses. Until that is done, the absolute values in the anchor's Table 3 cannot be read as fairness magnitudes, and the stable-output, high-shift quadrant cannot be distinguished from the stable-output, high-

shift quadrant a placebo would populate.

7. Three readings of a stable output with a moving interior

Suppose the nulls were in place and a population of users still showed large attribute-specific internal movement with no attribute-specific change in recommendations. The anchor names this silent internal bias. At least three mechanisms produce it, and they have opposite implications for fairness.

7.1 Encoded and declined

The model registers the attribute and does not let it affect the ranking. On this reading the fourth quadrant is where a well-behaved model lives. There is direct evidence that registering a protected attribute can be what allows a model to avoid acting on it. Sun et al. (2025) found that aligned Llama 3 models were less likely than their base counterparts to interpret the words black and white as racial in ambiguous contexts (52.2 percent less likely, by activation patching), that the aligned 70B model's implicit-association responses linked black with negativity, guilt and weapons 91.4 percent of the time, and that injecting race-related activations into Llama 3 8B Instruct cut implicit bias from 97.3 percent to 71.2 percent, with early-layer LoRA fine-tuning bringing it to 42.3 percent. Their interpretation is that the representation of race activates safeguards which, absent the representation, likely fail to engage. Ganguli et al. (2023) found that sufficiently large RLHF-trained models reduce stereotyped answers when instructed to, a capability they report emerging at 22B parameters, which presupposes that the model registers the group a question concerns. Tamkin et al. (2023) reduced both positive and negative discrimination in Claude 2.0's decisions through prompt interventions added while the demographic information stayed in the prompt. An early statement of the principle is Dwork et al. (2011), whose framework treats similar individuals similarly under a task-specific similarity metric and whose title names the alternative to ignoring the attribute: fairness pursued with awareness of it.

7.2 Used, but cancelled

The attribute flows to the decision through more than one pathway, and the effects cancel in the audited condition. Vig et al. (2020) found gender-bias effects in pre-trained Transformer language models to be sparse and synergistic, some components amplifying and others repressing them. A cancelled effect is fragile: a perturbation that weakens one pathway exposes the other. On this reading the fourth quadrant is a genuine risk, but the risk is not established by the internal shift; it is established by finding the perturbation that breaks the cancellation.

7.3 Latent, and returned by a perturbation

The attribute is not used now, but a stated perturbation would make the model use it. This is the reading closest to the anchor's concern, and it has the strongest published support of the three as a phenomenon, though not as an explanation of FairGap's fourth quadrant. Bouchouchi et al. (2026) found that LoRA fine-tuning against gender bias lowered expressed bias (one profession's score moved from -0.90 to -0.17) while latent polarisation stayed above the random-direction reference, and that a single jailbreak instruction restored expressed polarisation to base-model levels or beyond. Orgad et al. (2022) found that bias suppressed by extrinsic debiasing returned when the classifier head was retrained on the original data, by an amount predictable from extractability. Kaneko et al. (2022) found debiased masked models relearning bias under fine-tuning. Hofmann et al. (2024) found that human-feedback training left covert dialect prejudice in place and widened the gap between covert and overt stereotypes, and Bai et al. (2025) found explicitly unbiased models exhibiting implicit associations that predicted biased decisions. Outside fairness, Qi et al. (2023) found safety alignment degraded by fine-tuning on a handful of adversarial examples and, less, by benign data; Al Hakim et al. (2026) found quantisation degraded fairness benchmarks across languages; and Goel et al. (2026) found that information about supposedly unlearned data that remained linearly decodable correlated with susceptibility to reconstruction attacks.

7.4 What separates them

The three readings predict the same point in the IBS-OBS plane. They predict different outcomes under three tests. Under a mediation test, ablating the attribute direction with a random-direction control, reading 7.1 predicts that output disparity rises, because the safeguard loses its input; reading 7.2 predicts that it may move either way depending on which pathway the ablation weakens more; reading 7.3 predicts little change in the audited condition. Under a transport test, re-auditing outputs after a stated perturbation (an adversarial instruction, a different prompt family, a fine-tuning step, quantisation), reading 7.3 predicts that the users with the largest attribute-specific internal shift are the ones

whose outputs diverge most after the perturbation, reading 7.1 predicts no such concentration, and reading 7.2 predicts concentration only for perturbations that touch one pathway. Under an awareness test, strengthening rather than removing the attribute representation as Sun et al. (2025) did, reading 7.1 predicts reduced disparity and reading 7.3 does not.

Orgad et al. (2022) and Bouchouchi et al. (2026) have each run a version of the transport test, and in both the internal measure was informative about what came back. That is the use of a hidden measure the published record supports: as a forecast of output disparity under a named perturbation, scored by whether it predicts. FairGap's internal axis has not been scored that way.

8. Reading the erasure result

The anchor's intervention found that removing a gender direction raised output disparity in all three measured conditions, and concluded that internal and output fairness can pull in opposite directions. The concept-erasure literature predicts several reasons why this particular intervention would raise disparity, and only one of them supports the anchor's conclusion.

8.1 What a mean-difference projection removes

A projection along the difference of class means removes the single linear direction along which the two groups' average representations differ. It does not remove the attribute. Elazar and Goldberg (2018) found that demographic information an adversary could no longer detect during training could still be recovered by a classifier trained afterwards. Gonen and Goldberg (2019) found that debiased word embeddings still encoded gender in the geometry of neighbouring words. Ravfogel et al. (2020) needed iterated projections to approach linear guarding, and Ravfogel et al. (2022) showed that protection against one nonlinear adversary does not transfer to another. Kumar et al. (2022) proved that probe-based removal is likely to remove features other than the concept, and that it can destroy task-relevant information without removing the concept. LEACE (Belrose et al., 2023) gives a closed-form guarantee against all linear adversaries, which a single mean-difference direction does not. What remained after FairGap's projection, and what else was removed with it, was not measured.

8.2 The missing perturbation control

Any edit to the prefill activations of five layers is a perturbation, and greedy list generation is sensitive to perturbations (Section 6.1). Korznikov et al. (2025) found that adding random directions to a model's activations was enough to move harmful-compliance rates from zero to between 1 and 13 percent. Makelov et al. (2023) showed that a subspace intervention can change behaviour by activating a dormant pathway rather than by editing the feature it targets. The author's

earlier analysis of steering argued that an intervention's effect licenses a claim about the targeted direction only relative to comparison directions of matched size. FairGap applied no random-direction projection. If projecting out a random direction of the same norm at the same layers raised OBS by a similar amount, the result would say nothing about gender. The test is inexpensive and decisive.

8.3 Removing awareness

If reading 7.1 holds for these models, erasing the gender direction removes what the model needs to keep gender from affecting the ranking, and output disparity should rise. That is what was observed. On this account the result is not a tension between internal and output fairness. It is evidence that the internal shift the benchmark counts as bias is, at least in part, the representation that keeps outputs fair, which is the reverse of the anchor's interpretation. Sun et al. (2025) found the corresponding effect in the other direction: strengthening the race representation reduced implicit bias.

8.4 The counter-record

Other interventions on representations and weights have lowered output bias. Bouchouchi et al. (2026) removed a gender direction at every layer and token and expressed polarisation fell. Limisiewicz et al. (2023) located gender bias in mid-to-upper feed-forward layers and reduced it with a linear projection of their weights while keeping downstream performance. Chintam et al. (2023) used causal mediation, circuit discovery and differential masking to find GPT-2 components responsible for gender bias and fine-tuned them with less damage than full fine-tuning. Cai et al. (2024) traced gender bias in occupational pronouns to particular MLP and attention modules and edited them. Li et al. (2025) reported inference-time steering that reduced bias across question answering, counterfactual inputs and open generation in six models, Sharma et al. (2025) derived steering targets that guarantee group-fair outcomes for parametric families, and Paes et al. (2025) noted that existing steering methods struggle when equiprobable outcomes across groups are required and optimised the steering transformation directly.

The pattern across these results is that the sign of an intervention's effect on output disparity depends on what is removed, where, and what the outcome measures. Removing a direction everywhere and scoring which gender a completion assigns is a different experiment from removing a direction at five layers during prefill and scoring how much a recommendation list differs between two stated genders. The first targets the pathway by which gender reaches the output word. The second targets the difference between two prompts at the point where the model has just read them, which is where reading 7.1 would locate the model's knowledge of whom it is serving. That the two experiments have

opposite effects is what the four-property account predicts.

9. What the normative layer adds

Even with every null in place and every mechanism identified, a hidden-output audit of a recommender inherits a normative question that its measurements cannot settle: when should a stated identity change what is recommended?

Recommendation is a personalisation task, and the multi-stakeholder fairness literature has long noted that this complicates the transfer of classification fairness concepts (Burke, 2017; Ekstrand et al., 2021), and a survey of the area found its fairness definitions numerous and its research scattered (Wang et al., 2022). Deldjoo and Di Noia (2024) argue that counting every change in a list as bias ignores whether the change tracks the user's actual preferences, and propose scoring alignment with held-out preferences. Kantharuban et al. (2024) frame the question as stereotype or personalisation and argue that the remedy is disclosure: a system should say when identity moved its recommendations. Li et al. (2021) treat fairness in recommendation as a per-user demand that some users may make and others may not. On none of these views is counterfactual invariance of the list to a stated identity the correct target in every case.

The counterfactual itself is contested. Kusner et al. (2017) define counterfactual fairness by comparison with a world in which the individual belongs to another group, which requires a causal model of how group membership affects the other features; Kilbertus et al. (2017) argue that fairness criteria should be stated as assumptions about the causal process. Kasirzadeh and Smart (2021) and Hu and Kohler-Hausmann (2020) argue that categories such as race and sex do not admit the clean manipulation such counterfactuals require, because many of their supposed effects are part of what the category is. A recommender audit that holds a user's history fixed and swaps a stated identity sidesteps this partly, because it asks a narrower question: does the stated identity override the evidence of the history? That is the robustness question Garg et al. (2018) formalised as counterfactual token fairness. It is a reasonable question, but it is a question about the effect of a sentence, and an internal shift caused by a sentence is a fact about reading.

9.1 Stated identity and inferred identity

The minimal-pair design has a second limit that bears directly on the hidden axis. FairGap states the attribute in a sentence of its own. Most identity information reaching a deployed recommender is not stated; it is inferred, from a name, from the way a request is written, or from the history itself. Salinas et al. (2024) found that advice from several models, GPT-4 among them, systematically disadvantaged names associated with

racial minorities and with women, consistently across 42 prompt templates, with names associated with Black women faring worst, and that supplying a numerical anchor relevant to the decision counteracted the bias while qualitative detail had inconsistent effects and sometimes widened it. Xu et al. (2023) found LLM rankers treating users differently on the basis of non-sensitive profile features such as names, which they call implicit ranking unfairness, and argued that it is more widespread than the explicit kind. Hofmann et al. (2024) found that the dialect of a text alone moved models' hypothetical decisions about employability and criminality, with covert stereotypes more negative than any human stereotype about African Americans experimentally recorded. Chen et al. (2024) found that an open chat model maintains an internal model of its user from which age, gender, education and socioeconomic status can be read, and built an interface through which users could inspect and adjust that internal model and thereby change the system's behaviour.

Two consequences follow for hidden-output audits. First, under inferred identity the internal axis has a job the explicit design cannot give it. When the attribute is never written, there is no comprehension cost to subtract: a representation of the user's gender built from a name or a writing style is the model's inference, and its movement under a name swap is evidence of inference in a way that movement under a stated-gender swap is not. The internal measure is most informative exactly where the anchor did not apply it. Second, an explicit-cue audit tests whether the model lets a stated attribute override the history; it does not test whether the model infers the attribute from the history and acts on that. A model can pass the first and fail the second. Kantharuban et al. (2024) found stereotyped recommendations whether identity was revealed on purpose or through implicit cues, which suggests the two can co-occur, but no study located compares a model's explicit-cue and inferred-cue disparities on the same users with the internal axis measured on both.

This gives the inferred-identity version of the transport test a natural form. Swap names, or rewrite a request in a different dialect, holding stated content fixed; measure the signed projection of the hidden-state difference onto an attribute direction learned from explicit statements; and ask whether that projection predicts the output disparity across users. If it does, a hidden measure has done what an output audit cannot: identified which users the model has silently assigned to a group. That would be a hidden bias in the sense the anchor intends, and the published record does not yet contain the measurement.

9.2 Representations with a known consumer

The representation-learning tradition supplies one more distinction. Edwards and Storkey (2015) and Madras et al. (2018) argued for removing protected information

from representations so that any downstream user of the representation acts fairly; Madras et al. framed the goal as fairness that transfers to tasks the representation's designer does not know. Lechner et al. (2021) proved that no representation can guarantee fairness for all downstream tasks, and that even demographic parity fails once the input distribution shifts. The intermediate state of a decoder-only recommender is not a released representation: its only consumer is the rest of the same fixed model. A representation-level fairness criterion has a rationale when the downstream user is unknown. When the downstream computation is known and fixed, the question is what that computation does, which is a question about use and expression, not about encoding.

10. A licensing procedure

Algorithm 1 states what a hidden-output audit of a recommender may conclude, given which tests were run. It is a decision structure over published test types, not a new metric.

Applied to the published record, the procedure gives the following. FairGap reaches step 2 without its null and stops at step 3, because its internal quantity is a distance; its steering experiment is a version of step 4 without the random control. Bouchouchi et al. (2026) reach step 4 with a random-direction reference for encoding though not for ablation, and a version of step 6 for one perturbation (a jailbreak instruction), though they test reactivation in aggregate rather than whether per-item internal scores forecast per-item output change. Orgad et al. (2022) reach a version of step 6, at the level of models rather than items, for one perturbation (head retraining). Sun et al. (2025) run the awareness test of step 5 for implicit association prompts. No study located runs all steps on a recommender.

11. Measurements that would settle the open part

Six studies would decide what the hidden-output gap is.

M1. Placebo-anchored FairGap. Re-run the benchmark with attribute-free and paraphrase swaps for every user, and report OBS, IBS and ROA on the excess over the nulls. If the excess IBS is near zero, the internal axis measured reading, not bias. If the excess IBS is substantial and its quadrant shares depart from independence, the fourth quadrant becomes a finding.

M2. Directed internal quantity. Replace cosine distance with the signed projection of the hidden-state difference onto an attribute direction estimated on held-out pairs, with a random-direction reference as in Bouchouchi et al. (2026), and with standardised dimensions as Timkey and van Schijndel (2021) recommend. Report ROA under all three extraction rules.

M3. Controlled ablation. Repeat the projection experiment with matched random directions at the same

Algorithm 1. Licensing a claim from a hidden-output fairness audit.

```

Input: model M, minimal pairs P differing in attribute A,
       placebo pairs P0 (attribute-free swap) and P1 (paraphrase of A),
       readout R on outputs, internal quantity I, perturbation class K
1  Compute R and I on P, P0, P1. Let  $dR = R(P) - \max(R(P0), R(P1))$ ;
   let  $dI = I(P) - \max(I(P0), I(P1))$ .
2  If  $dR$  is not above its null: no expressed disparity under R.
   Else: expressed disparity licensed, at magnitude  $dR$ , for R only.
3  If I is undirected (a distance): report  $dI$  as exposure, not bias; stop
   unless step 4 is run with an A-specific direction.
4  Estimate direction  $d_A$ ; compare ablation of  $d_A$  with ablation of
   random directions of matched norm at the same sites.
   If effect( $d_A$ ) does not exceed the random spread: A not shown used.
   If it does: A used for R at those sites; record the sign.
5  If sign is (disparity rises when  $d_A$  removed): run awareness test
   (strengthen  $d_A$ ); if disparity falls, classify as encoded-and-declined.
6  For each perturbation k in K: re-run step 2 under k; test whether
   per-item  $dI$  forecasts per-item  $dR$  under k (rank correlation, held out).
   If yes for k: A is latent for R under k. Claim nothing for k not in K.
7  Report the licensed claims with their R, I, sites and K; report no
   claim of hidden bias that steps 2 to 6 did not establish.

```

layers, with a LEACE projection (Belrose et al., 2023), and with removal at all layers and tokens. If random projections raise OBS as much as the gender projection does, the steering result reflects generic perturbation sensitivity.

M4. Awareness intervention. Add, rather than remove, the attribute direction at early layers, as Sun et al. (2025) did for race. If output disparity falls, the encoded-and-declined reading holds for those models and the fourth quadrant is evidence of fairness, not of its absence.

M5. Transport forecast. Choose a perturbation class in advance (an adversarial instruction asking for demographically typical recommendations, a fine-tuning step on interaction logs, 4-bit quantisation) and test, on held-out users, whether pre-perturbation internal shift predicts post-perturbation output shift. This is the test that would turn a hidden measure into an audit instrument.

M6. Inferred-identity audit. Repeat M1 to M5 with identity carried by names or dialect instead of a stated sentence, as in the designs of Salinas et al. (2024) and Hofmann et al. (2024), on the same users. Compare explicit-cue and inferred-cue output disparities, and ask whether the internal projection predicts which users the model treats as members of a group when nobody told it. Section 9.1 argues this is where an internal measure has the most to add.

12. Objections

"The anchor already reports robustness checks." It does, for the persistence parameter of rank-biased overlap, for thresholds, for prompt families and for age phrasing. None of them is a placebo, and the extraction check, the one that changes the internal quantity, reverses the sign of the central statistic. Robustness to choices on the output side does not establish validity on the internal side.

"Any hidden movement under an identity swap is a fairness concern, whatever its mechanism." That position makes every model that reads its input unfair, since a model must represent a stated attribute in order to parse the sentence that states it. It also makes the remedy of Section 7.1, attending to the attribute in order not to act on it, count as a failure. A fairness criterion that a model satisfies only by failing to understand its prompt is not one that audits should adopt.

"Bouchouchi et al. show that encoded bias matters, so the hidden axis is vindicated." Their result vindicates a different internal quantity, a signed projection on an attribute direction, measured across entities against a random reference and backed by ablation. It is evidence that the right internal measure can be informative, which is the paper's position, not evidence that cosine distance is that measure.

"The paper relies on derivations, not measurements." The quadrant comparison in Section 4.4 and Table 3 are derivations and are labelled as such. The derivation in Section 4.4 is arithmetic on the anchor's printed shares, and anyone can check it. The predictions in Table 3 follow from the readings as defined, and Section 11 says how each would be tested. Every quantitative statement in the paper is either a value printed by a cited source or arithmetic on such values, and each value taken from the anchor is given with the anchor table that prints it.

"One anchor is a thin base for a general argument." The paper's general claims rest on the encoded-versus-expressed literature (Section 5), the concept-erasure literature (Section 8) and the probing literature (Section 3), not on FairGap alone. FairGap is the occasion and the clearest instance of an undirected internal axis proposed as a fairness measure. If later versions of the benchmark add placebos and a directed quantity, much of Section 4 will no longer apply, and that is the intended outcome.

13. What is not known

It is not known how large the attribute-free placebo shift is, on either axis, for any of the six FairGap models; no source reports it. It is not known whether the gender projection's effect on output shift exceeds that of a random projection of matched norm at the same layers. It is not known whether any recommender's internal shift under an identity swap forecasts its output disparity under a perturbation the audit did not run. It is not known whether the encoded-and-declined mechanism found for race in implicit association prompts (Sun et al., 2025) operates in recommendation. It is not known whether the strong within-model correlation found across entities for gendered concepts (Bouchouchi et al., 2026) holds across users in a recommender. And it is not known which, if any, users in FairGap's fourth quadrant receive worse recommendations, since neither axis scores harm to a user.

14. Conclusion

The hidden-output fairness gap was proposed as evidence that output-only audits of language-model recommenders are incomplete. The published evidence supports a narrower and more useful conclusion. What a model encodes about a protected attribute is related to what it expresses when the internal measure is tied to the attribute and the two measures share their inputs; the older finding that the two are unrelated appears to have been mostly a finding about mismatched instruments. An undirected distance between hidden states, by contrast, measures that the model read a different sentence. Its correlation with output shift depends on how it is summarised, in the one cell where the anchor reports percentile thresholds its quadrant shares are what independence gives, and the intervention that lowered it and raised output disparity is at least as well explained by the removal of the representation a model uses to stay fair as by a hidden bias. The stable-output, moving-interior case is where a fair model and a latently biased one both sit, and only a mediation test and a transport test can tell them apart. An internal measure earns a place in a fairness audit when it forecasts output disparity under a perturbation stated in advance. None of the evidence reviewed here shows that the undirected distance does.

References

- Al Hakim, M. A., Wicaksono, A. F., & Koto, F. (2026). Preserving Fairness and Safety in Quantized LLMs Through Critical Weight Protection. arXiv:2601.12033. doi:10.48550/arXiv.2601.12033. [PREPRINT]
- Arzaghi, M., Farashah, A. D., Carichon, F., Plante, J.-F., & Farnadi, G. (2025). Let's Unlearn Stereotypes Before Decision-Making: Assessing the Impact of Intrinsic Bias Mitigation on Downstream Fairness in LLMs. arXiv:2509.16462. doi:10.48550/arXiv.2509.16462. [PREPRINT]
- Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., et al. (2024). Non-Determinism of "Deterministic" LLM Settings. arXiv:2408.04667. doi:10.48550/arXiv.2408.04667.
- [PREPRINT]
- Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly Unbiased Large Language Models Still Form Biased Associations. Proceedings of the National Academy of Sciences, 122(8). arXiv:2402.04105 (as "Measuring Implicit Bias in Explicitly Unbiased Large Language Models"). doi:10.48550/arXiv.2402.04105.
- Belinkov, Y. (2021). Probing Classifiers: Promises, Shortcomings, and Advances. arXiv:2102.12452. doi:10.48550/arXiv.2102.12452.
- Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2023). LEACE: Perfect linear concept erasure in closed form. arXiv:2306.03819. doi:10.48550/arXiv.2306.03819.
- Bouchouchi, N., Laugel, T., Renard, X., Marsala, C., Lesot, M.-J., & Detryniecki, M. (2026). Alignment Reduces Expressed but Not Encoded Gender Bias: A Unified Framework and Study. arXiv:2603.24125. doi:10.48550/arXiv.2603.24125. [PREPRINT]
- Burke, R. (2017). Multisided Fairness for Recommendation. arXiv:1707.00093. doi:10.48550/arXiv.1707.00093.
- Cai, Y., Cao, D., Guo, R., Wen, Y., Liu, G., & Chen, E. (2024). Locating and Mitigating Gender Bias in Large Language Models. arXiv:2403.14409. doi:10.48550/arXiv.2403.14409. [PREPRINT]
- Cao, Y. T., Pruksachatkun, Y., Chang, K.-W., Gupta, R., Kumar, V., Dhamala, J., & Galstyan, A. (2022). On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. arXiv:2203.13928. doi:10.48550/arXiv.2203.13928.
- Chen, Y., Wu, A., DePodesta, T., Yeh, C., Li, K., Marin, N. C., et al. (2024). Designing a Dashboard for Transparency and Control of Conversational AI. arXiv:2406.07882. doi:10.48550/arXiv.2406.07882. [PREPRINT]
- Chintam, A., Beloch, R., Zuidema, W., Hanna, M., & van der Wal, O. (2023). Identifying and Adapting Transformer-Components Responsible for Gender Bias in an English Language Model. arXiv:2310.12611. doi:10.48550/arXiv.2310.12611.
- Deldjoo, Y., & Di Noia, T. (2024). CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System. arXiv:2403.05668. doi:10.48550/arXiv.2403.05668. [PREPRINT]
- Delobelle, P., Tokpo, E. K., Calders, T., & Berendt, B. (2021). Measuring Fairness with Biased Rulers: A Survey on Quantifying Biases in Pretrained Language Models. arXiv:2112.07447. doi:10.48550/arXiv.2112.07447.
- Dwivedi-Yu, J., Dwivedi, R., & Schick, T. (2024). FairPair: A Robust Evaluation of Biases in Language Models through Paired Perturbations. arXiv:2404.06619. doi:10.48550/arXiv.2404.06619. [PREPRINT]
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. arXiv:1104.3913. doi:10.48550/arXiv.1104.3913.
- Edwards, H., & Storkey, A. (2015). Censoring Representations with an Adversary. arXiv:1511.05897. doi:10.48550/arXiv.1511.05897.
- Ekstrand, M. D., Das, A., Burke, R., & Diaz, F. (2021). Fairness in Information Access Systems. arXiv:2105.05779. doi:10.48550/arXiv.2105.05779.
- Elazar, Y., & Goldberg, Y. (2018). Adversarial Removal of Demographic Attributes from Text Data. arXiv:1808.06640. doi:10.48550/arXiv.1808.06640.
- Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2020). Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. arXiv:2006.00995. doi:10.48550/arXiv.2006.00995.
- Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. arXiv:1909.00512. doi:10.48550/arXiv.1909.00512.
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., et al. (2024). Bias and Fairness in Large Language Models: A Survey. Computational Linguistics, 50(3).

- arXiv:2309.00770. doi:10.48550/arXiv.2309.00770.
- Ganguli, D., Askell, A., Schiefer, N., Liao, T. I., Lukosiute, K., Chen, A., et al. (2023). The Capacity for Moral Self-Correction in Large Language Models. arXiv:2302.07459. doi:10.48550/arXiv.2302.07459. [PREPRINT]
- Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., & Beutel, A. (2018). Counterfactual Fairness in Text Classification through Robustness. arXiv:1809.10610. doi:10.48550/arXiv.1809.10610.
- Goel, A., Ritter, A., & Gurevych, I. (2026). Auditing Language Model Unlearning via Information Decomposition. arXiv:2601.15111. doi:10.48550/arXiv.2601.15111. [PREPRINT]
- Goldfarb-Tarrant, S., Marchant, R., Sanchez, R. M., Pandya, M., & Lopez, A. (2021). Intrinsic Bias Metrics Do Not Correlate with Application Bias. arXiv:2012.15859. doi:10.48550/arXiv.2012.15859.
- Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. arXiv:1903.03862. doi:10.48550/arXiv.1903.03862.
- Hewitt, J., & Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. arXiv:1909.03368. doi:10.48550/arXiv.1909.03368.
- Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv:2403.00742. doi:10.48550/arXiv.2403.00742.
- Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., & Zhao, W. X. (2023). Large Language Models are Zero-Shot Rankers for Recommender Systems. arXiv:2305.08845. doi:10.48550/arXiv.2305.08845.
- Hu, L., & Kohler-Hausmann, I. (2020). What's Sex Got To Do With Fair Machine Learning? arXiv:2006.01770. doi:10.48550/arXiv.2006.01770.
- Hua, W., Ge, Y., Xu, S., Ji, J., & Zhang, Y. (2024). UP5: Unbiased Foundation Model for Fairness-aware Recommendation. EACL 2024. arXiv:2305.12090. doi:10.48550/arXiv.2305.12090.
- Jin, X., Barbieri, F., Kennedy, B., Davani, A. M., Neves, L., & Ren, X. (2021). On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning. arXiv:2010.12864. doi:10.48550/arXiv.2010.12864.
- Kaneko, M., Bollegala, D., & Okazaki, N. (2022). Debiasing isn't enough! – On the Effectiveness of Debiasing MLMs and their Social Biases in Downstream Tasks. arXiv:2210.02938. doi:10.48550/arXiv.2210.02938.
- Kantharuban, A., Milbauer, J., Sap, M., Strubell, E., & Neubig, G. (2024). Stereotype or Personalization? User Identity Biases Chatbot Recommendations. arXiv:2410.05613. doi:10.48550/arXiv.2410.05613.
- Kasirzadeh, A., & Smart, A. (2021). The Use and Misuse of Counterfactuals in Ethical Machine Learning. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. doi:10.1145/3442188.3445886.
- Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Scholkopf, B. (2017). Avoiding Discrimination through Causal Reasoning. arXiv:1706.02744. doi:10.48550/arXiv.1706.02744.
- Korznikov, A., Galichin, A., Dontsov, A., Rogov, O. Y., Oseledets, I., & Tutubalina, E. (2025). The Rogue Scalpel: Activation Steering Compromises LLM Safety. arXiv:2509.22067. doi:10.48550/arXiv.2509.22067. [PREPRINT]
- Kumar, A., Tan, C., & Sharma, A. (2022). Probing Classifiers are Unreliable for Concept Removal and Detection. arXiv:2207.04153. doi:10.48550/arXiv.2207.04153.
- Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual Fairness. arXiv:1703.06856. doi:10.48550/arXiv.1703.06856.
- Lechner, T., Ben-David, S., Agarwal, S., & Ananthakrishnan, N. (2021). Impossibility results for fair representations. arXiv:2107.03483. doi:10.48550/arXiv.2107.03483. [PREPRINT]
- Li, Y., Chen, H., Xu, S., Ge, Y., & Zhang, Y. (2021). Personalized Counterfactual Fairness in Recommendation. arXiv:2105.09829. doi:10.48550/arXiv.2105.09829.
- Li, Y., Fan, Z., Chen, R., Gai, X., Gong, L., Zhang, Y., & Liu, Z. (2025). FairSteer: Inference Time Debiasing for LLMs with Dynamic Activation Steering. arXiv:2504.14492. doi:10.48550/arXiv.2504.14492.
- Limisiewicz, T., Marecek, D., & Musil, T. (2023). Debiasing Algorithm through Model Adaptation. arXiv:2310.18913. doi:10.48550/arXiv.2310.18913.
- Lu, C. A., Fayyazi, A., Zhang, J., Shokoufa, S., Xing, Y., Xiang, Z., et al. (2026). Fair on the Surface? Benchmarking Hidden-Output Fairness Gaps in LLM Recommenders. arXiv:2608.08284. doi:10.48550/arXiv.2608.08284. [PREPRINT]
- Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning Adversarially Fair and Transferable Representations. arXiv:1802.06309. doi:10.48550/arXiv.1802.06309.
- Makelov, A., Lange, G., & Nanda, N. (2023). Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. arXiv:2311.17030. doi:10.48550/arXiv.2311.17030.
- Orgad, H., Goldfarb-Tarrant, S., & Belinkov, Y. (2022). How Gender Debiasing Affects Internal Model Representations, and Why It Matters. arXiv:2204.06827. doi:10.48550/arXiv.2204.06827.
- Paes, L. M., Sivakumar, N., Wang, Y. O., Fedzechkina, M., Theobald, B.-J., Zappella, L., & Apostoloff, N. (2025). DSO: Direct Steering Optimization for Bias Mitigation. arXiv:2512.15926. doi:10.48550/arXiv.2512.15926. [PREPRINT]
- Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv:2310.03693. doi:10.48550/arXiv.2310.03693.
- Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. arXiv:2004.07667. doi:10.48550/arXiv.2004.07667.
- Ravfogel, S., Vargas, F., Goldberg, Y., & Cotterell, R. (2022). Kernelized Concept Erasure. arXiv:2201.12191. doi:10.48550/arXiv.2201.12191.
- Ravichander, A., Belinkov, Y., & Hovy, E. (2020). Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? arXiv:2005.00719. doi:10.48550/arXiv.2005.00719.
- Salinas, A., Haim, A., & Nyarko, J. (2024). What's in a Name? Auditing Large Language Models for Race and Gender Bias. arXiv:2402.14875. doi:10.48550/arXiv.2402.14875. [PREPRINT]
- Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. arXiv:2310.11324. doi:10.48550/arXiv.2310.11324.
- Seshadri, P., Pezeshkpour, P., & Singh, S. (2022). Quantifying Social Biases Using Templates is Unreliable. arXiv:2210.04337. doi:10.48550/arXiv.2210.04337. [PREPRINT]
- Sharma, M., Deshpande, A. J., Bhattacharyya, C., & Shah, R. R. (2025). On Optimal Steering to Achieve Exact Fairness. arXiv:2509.15759. doi:10.48550/arXiv.2509.15759. [PREPRINT]
- Steed, R., Panda, S., Kobren, A., & Wick, M. (2022). Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). doi:10.18653/v1/2022.acl-long.247.
- Sun, L., Mao, C., Hofmann, V., & Bai, X. (2025). Aligned but Blind: Alignment Increases Implicit Bias by Reducing Awareness of Race. ACL 2025. arXiv:2506.00253. doi:10.48550/arXiv.2506.00253.
- Sun, M., Chen, X., Kolter, J. Z., & Liu, Z. (2024). Massive Activations in Large Language Models. arXiv:2402.17762. doi:10.48550/

- arXiv.2402.17762.
- Tamkin, A., Aspell, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., et al. (2023). Evaluating and Mitigating Discrimination in Language Model Decisions. arXiv:2312.03689. doi:10.48550/arXiv.2312.03689. [PREPRINT]
- Timkey, W., & van Schijndel, M. (2021). All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality. arXiv:2109.04404. doi:10.48550/arXiv.2109.04404.
- Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., et al. (2020). Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. arXiv:2004.12265. doi:10.48550/arXiv.2004.12265.
- Wang, Y., Ma, W., Zhang, M., Liu, Y., & Ma, S. (2022). A Survey on the Fairness of Recommender Systems. arXiv:2206.03761. doi:10.48550/arXiv.2206.03761.
- Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. ACM Transactions on Information Systems, 28(4). doi:10.1145/1852102.1852106.
- Xu, C., Wang, W., Li, Y., Pang, L., Xu, J., & Chua, T.-S. (2023). A Study of Implicit Ranking Unfairness in Large Language Models. arXiv:2311.07054. doi:10.48550/arXiv.2311.07054.
- Zhang, J., Bao, K., Zhang, Y., Wang, W., Feng, F., & He, X. (2023). Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. RecSys 2023. arXiv:2305.07609. doi:10.48550/arXiv.2305.07609.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23437257