

Fair on the Surface? Benchmarking Hidden–Output Fairness Gaps in LLM Recommenders

Chan Aristella Lu*
University of Georgia

Arya Fayyazi*
University of Southern California

Junhao Zhang
Carnegie Mellon University

Saeid Shokoufa
University of Southern California

Yue Xing
Michigan State University

Zhen Xiang
University of Georgia

Kyu Hyung Lee
University of Georgia

Mehdi Kamal
University of Southern California

Massoud Pedram
University of Southern California

Abstract

Fairness audits for LLM-based recommenders have largely focused on observable outputs, implicitly assuming that stable recommendations reflect stable internal processing. We challenge this assumption with FAIRGAP, the first benchmark to jointly evaluate recommendation fairness at two levels: observable output shift (OBS) and hidden representation shift (IBS), measured through controlled counterfactual identity probes across gender, age, and race. Their relationship is summarized via Representation–Output Alignment (ROA), with quadrant diagnostics for identifying user-level hidden–output mismatch. Applied to six open-weight LLM families across three domains, FAIRGAP reveals pervasive hidden–output decoupling: ROA rarely exceeds 0.22, and a non-negligible user population shows stable outputs despite substantial internal shifts, a mode that output-only audits cannot detect by design. Further, activation steering that reduces IBS by up to $8\times$ simultaneously worsens OBS, demonstrating a fundamental tension between internal and output-level fairness that existing frameworks are unequipped to diagnose.

1 Introduction

Large language models (LLMs) are increasingly used as recommenders in conversational agents [1, 2], search platforms [3, 4], and personalized recommendation interfaces [5, 6]. This growing use has raised important fairness concerns, as recommendation outcomes can differ across user attributes such as gender, age, or race even when the underlying preference history is held fixed [7, 8]. Prior work has identified these disparities and proposed mitigation methods [7, 8], but the evaluations used to assess them remain output-centric and overlook representational changes inside the model [9–11]. Consequently, recommendation outputs that appear fair on the surface can still arise from substantially different internal representations [12–14], making output-only evaluation insufficient for assessing fairness or determining the reliability of mitigation over time [15, 16].

A key limitation of output-only fairness evaluation is that it treats the recommendation outputs as the sole evidence of fairness. For example, widely used benchmarks such as WinoBias for biased coreference [17], CrowS-Pairs for stereotype likelihood [18], BBQ for biased question answering [19], and RealToxicityPrompts for toxic generation [20] evaluate fairness through outputs rather than the internal representations that produce them. Yet in LLM recommenders, the recommendation outputs are only the observable outcome of a latent inference process that standard audits cannot directly inspect. Therefore, fairness at the surface level may fail to reflect how the model internally

*These authors contributed equally. Contact: c125054@uga.edu, afayyazi@usc.edu

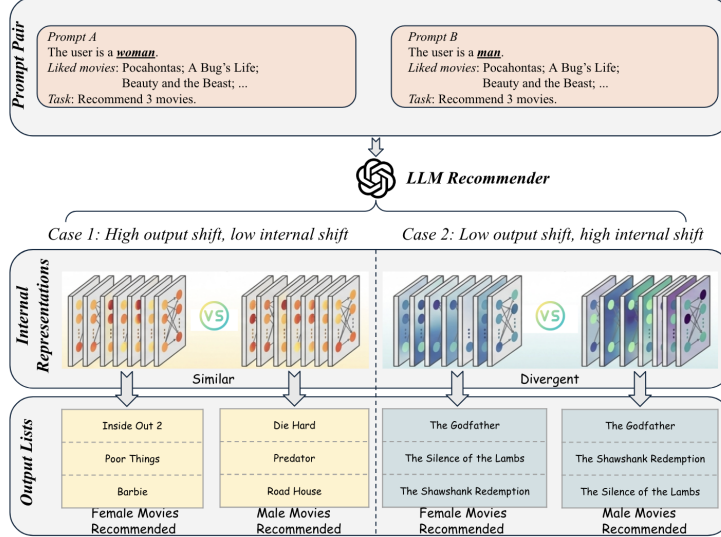


Figure 1: **Motivation for FairGap.** Two counterfactual recommendation prompts keep the same preference history and task, changing only the gender cue. Output-only fairness evaluation cannot distinguish between two qualitatively different cases: one in which recommendation outputs diverge despite similar internal representations, and another in which outputs remain similar enough to appear fair on the surface despite marked internal divergence.

processes protected-attribute cues [12–14]. Two counterfactual prompts can produce markedly different recommendation outputs even when the model’s internal representations remain similar, or they can produce highly similar outputs while relying on sharply different internal representations. Figure 1 illustrates this hidden–output fairness gap. In Case 1, changing only the gender cue produces a clear shift in the movie recommendation lists despite relatively similar internal representations. In Case 2, the movie recommendation lists remain similar and may appear fair on the surface, even though the internal representations diverge. These contrasting cases show that output-level behavior alone is insufficient for fairness diagnosis: visible disparities do not necessarily indicate major internal change, while output similarity can still mask fairness-sensitive internal divergence.

We introduce FAIRGAP, a benchmark that jointly measures *output shift* (OBS, rank-biased overlap on counterfactual ranked lists) and *internal shift* (IBS, cosine distance on mid-layer hidden states) across controlled demographic probes. Their alignment is captured by a Spearman correlation (ROA) and a four-quadrant taxonomy over the joint IBS–OBS space. As the sharpest diagnostic test of this design, activation steering reduces IBS by up to 8× while *simultaneously worsening* OBS in every tested condition, which reveals a failure mode that is entirely invisible to output-only auditing (Section 5.4).

Contributions. (1) We identify the *hidden–output fairness gap* as a distinct, previously unmeasured evaluation problem in LLM recommenders. (2) We release FAIRGAP, to our knowledge the first benchmark based on controlled counterfactual identity probes that jointly quantifies OBS and IBS across six open-weight LLM families, three recommendation domains, and three protected attributes. (3) We propose a four-quadrant IBS–OBS taxonomy enabling finer-grained, diagnostically interpretable fairness analysis. (4) We demonstrate through activation-steering experiments that internal and output-level fairness can diverge under intervention, establishing that joint evaluation is necessary for diagnosing hidden–output fairness gaps.

2 Related Work

Fairness in recommender systems. Recommendation fairness is inherently multi-sided, balancing user-side utility, item-side exposure, and aggregate quality [21, 22]. Group and individual fairness need not align because group parity can mask within-group disparities [23, 24], and even envy-freeness is insufficient in personalized settings [25]. As LLMs enter recommendation pipelines, these concerns extend to generative, free-form ranked outputs: prior work shows that sensitive user attributes shape LLM recommendations [26, 8], and benchmarks such as FaiRLLM operationalize fairness via demographic counterfactual output similarity [7]. FAIRGAP inherits this counterfactual framing but augments it with an internal representation axis.

Fairness evaluation in LLMs. LLM fairness is now studied across prediction disparities [9, 10], generation-based harms [27, 28], and counterfactual evaluation methods [29]. Evaluation has become predominantly output-centered: prompt-based audits [30, 11], uncertainty-aware metrics such as UCerF on SynthBias [31], and multi-turn benchmarks such as FairMT-Bench [32] all measure whether demographic variation changes externally observable behavior. The relationship between output-level change and internal representation change remains unresolved, and FAIRGAP is designed to address this gap.

Internal signals and representations. A complementary literature argues that model behavior cannot be read from outputs alone. Probing and representation-engineering work shows that sensitive attributes can remain nonlinearly encoded even after output-level debiasing [33, 34], and that fairness-oriented prompts can reduce stereotypical outputs while leaving biased continuations more probable across the layer stack [14]. Mechanistic interpretability studies use activation patching and attribution [35–37] to show that internal signals carry information invisible to output audits, including knowledge that persists linearly decodable even when output-based unlearning metrics report success [15]. Hernandez et al. further show that internal encodings can both steer and diagnose generation [16]. FAIRGAP operationalizes these insights at benchmark scale: we measure how minimal demographic perturbations propagate through hidden representations (IBS) jointly with their observable recommendation effects (OBS), enabling systematic diagnosis across model families, domains, and protected attributes.

Table 1: Comparison of representative fairness benchmarks. $\triangle$ denotes partial applicability to modern LLM evaluation.

Benchmark	Task Form	Rec. Task	LLM Eval.	Identity Cues	Output Fairness	Hidden–Output Gap
WinoBias [17]	Coreference	$\times$	$\times$	$\checkmark$	$\checkmark$	$\times$
CrowS-Pairs [18]	Sentence pairs	$\times$	$\times$	$\checkmark$	$\checkmark$	$\times$
BBQ [19]	Question answering	$\times$	$\triangle$	$\checkmark$	$\checkmark$	$\times$
RealToxicityPrompts [20]	Toxic generation	$\times$	$\triangle$	$\times$	$\checkmark$	$\times$
SynthBias [31]	Synthetic sentence/prompt	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$
FairEval [7]	LLM recommendation	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$
FairGap	LLM recommendation	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$

3 FairGap Framework

3.1 Problem Formulation

We study fairness in LLM recommenders under *counterfactual identity probes*. Each probe consists of a matched pair of prompts that share the same user preference history and task instruction while differing only in one protected-attribute cue, such as gender, age, or race. Let $x_u^{(a)}$ and $x_u^{(b)}$ denote the two prompts for user u , where the only difference is the protected-attribute value. Given a recommender model f , each prompt produces both a ranked recommendation list and the internal representation underlying that recommendation process.

Our goal is to quantify the extent to which the model changes when only the protected-attribute cue varies. We define the hidden–output fairness gap as the potential mismatch between the *visible* change in recommendation behavior and the *hidden* change in the model’s internal representations. Under the FairGap perspective, counterfactual stability corresponds to consistency on both dimensions while the preference history is held fixed. By contrast, output-only evaluation cannot detect cases where outputs remain stable despite substantial internal shifts, or where outputs change without commensurate internal movement. FairGap addresses this problem by measuring fairness jointly along two complementary axes: output shift and internal shift.

3.2 Output and Internal Fairness Axes

For each user-level counterfactual pair, we operationalize these two axes as an output-side distance and an internal-side distance. Let $Y_u^{(a)}$ and $Y_u^{(b)}$ denote the top- K recommendation lists produced for user u under the two protected-attribute conditions. We define the *output shift* as

$$d_{\text{out}}(u) = 1 - \text{RBO@}K(Y_u^{(a)}, Y_u^{(b)}),$$

where $\text{RBO@}K$ is the rank-biased overlap between the two recommendation lists [38], computed with persistence parameter p (Appendix A.1). This measure captures both item overlap and rank sensitivity, making it suitable for recommendations.

To measure hidden sensitivity, we extract hidden representations at four relative layer positions,

$$\mathcal{L} = \left\{ \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{4}{4} \right\},$$

which makes the internal measurement comparable across models with different numbers of layers. For a model with M layers, each relative depth $\ell \in \mathcal{L}$ is mapped to the nearest corresponding layer index. Let $h_{u,\ell}^{(a)}$ and $h_{u,\ell}^{(b)}$ denote the hidden representations at relative depth ℓ under the two protected-attribute conditions. We compute the layerwise internal shift as

$$\delta_{u,\ell} = 1 - \cos(h_{u,\ell}^{(a)}, h_{u,\ell}^{(b)}).$$

To obtain a single internal-shift score, we aggregate these four layerwise distances using weights proportional to protected-attribute separability:

$$d_{\text{in}}(u) = \sum_{\ell \in \mathcal{L}} \frac{\text{sep}_{\ell}}{\sum_{\ell' \in \mathcal{L}} \text{sep}_{\ell'}} \delta_{u,\ell},$$

where sep_{ℓ} denotes the protected-attribute separability at relative depth ℓ , defined from the development-set AUC of a linear probe trained to predict the protected attribute from hidden representations at that depth (Appendix A.2).

At the pair level, d_{out} and d_{in} characterize output-side and internal-side change, respectively. At the condition level, we summarize their relationship using *Representation–Output Alignment* (ROA):

$$\text{ROA} = \rho_{\text{Spearman}}(\{d_{\text{in}}(u)\}_{u=1}^N, \{d_{\text{out}}(u)\}_{u=1}^N),$$

where N is the number of matched counterfactual pairs in a given model–dataset–attribute condition in the evaluation split. ROA captures the extent to which internal change is reflected in observable recommendation change.

3.3 Quadrant Taxonomy

FairGap organizes user-level counterfactual pairs into a four-quadrant taxonomy over the joint space of $(d_{\text{out}}, d_{\text{in}})$ (Figure 2). **Q1** captures pairs with high output shift but low internal shift, indicating behavioral change without correspondingly large representational movement. **Q2** captures pairs with both high output shift and high internal shift, indicating coupled change across internal and output spaces. **Q3** captures pairs with low output shift and low internal shift, indicating stability on both dimensions. **Q4** captures pairs with low output shift but high internal shift, revealing latent internal perturbations that remain masked at the output level.

For each matched counterfactual pair u , FairGap computes a point in the joint shift space,

$$(d_{\text{out}}(u), d_{\text{in}}(u)),$$

and then estimates a threshold pair (t_x^*, t_y^*) using a two-dimensional Joint Otsu procedure [39](Appendix A.4). This procedure derives thresholds adaptively from the empirical joint distribution by selecting the partition that maximizes between-class scatter across the four induced regions. Each user pair is then assigned according to whether its output shift lies below or above t_x^* and whether its internal shift lies below or above t_y^* :

$$\begin{aligned} \text{Q1} : d_{\text{out}} \geq t_x^*, d_{\text{in}} < t_y^*, & \quad \text{Q2} : d_{\text{out}} \geq t_x^*, d_{\text{in}} \geq t_y^*, \\ \text{Q3} : d_{\text{out}} < t_x^*, d_{\text{in}} < t_y^*, & \quad \text{Q4} : d_{\text{out}} < t_x^*, d_{\text{in}} \geq t_y^*. \end{aligned}$$

This taxonomy provides a diagnostic view of fairness by distinguishing cases in which internal and output-level shifts move together from those in which they diverge, thereby making hidden–output mismatch directly observable. Because we treat the thresholds as diagnostic rather than normative, Appendix C.4 reports percentile-based alternatives as a robustness check.

3.4 Benchmark Usage

FAIRGAP is a reusable benchmark for evaluating hidden–output fairness gaps in LLM-based recommendation. A benchmark instance consists of a model, a user-profile dataset, and a protected attribute with counterfactual instantiations. For each user, matched counterfactual prompts are constructed by varying only the protected-attribute cue. The model produces top- K recommendations and hidden representations, from which FAIRGAP computes three metrics: OBS (output divergence), IBS (internal divergence), and ROA (their alignment). We recommend reporting OBS, IBS, and ROA jointly, with quadrant analysis providing a diagnostic view of hidden–output mismatch.

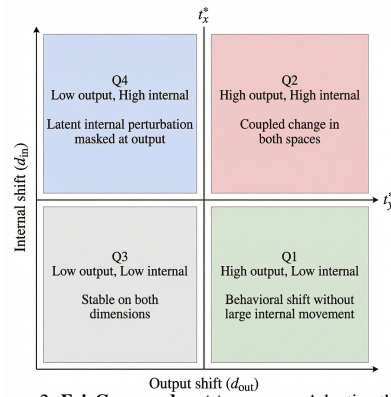


Figure 2: **FairGap quadrant taxonomy.** Adaptive thresholds (t_x^*, t_y^*) partition the joint space of output shift (d_{out}) and internal shift (d_{in}) into four diagnostic regions.

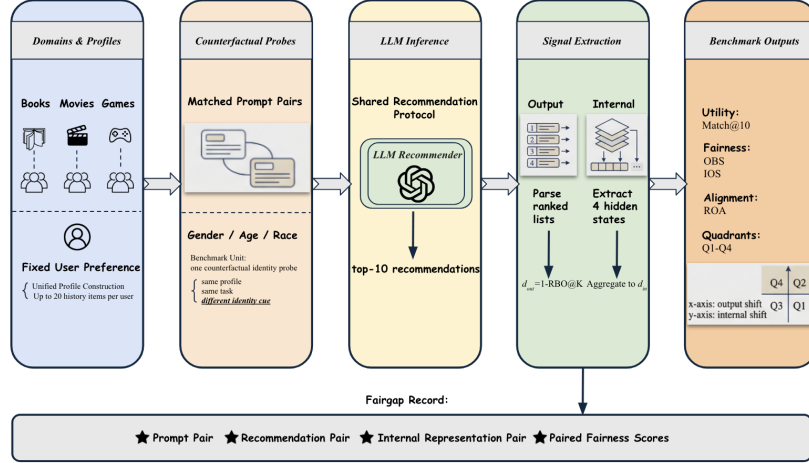


Figure 3: **FairGap benchmark construction pipeline.** Fixed user profiles are converted into matched counterfactual prompt pairs that differ only in a protected-attribute cue. These pairs are evaluated with a shared LLM recommendation protocol to produce top-10 recommendation lists and internal representations, which are then used to compute utility, fairness, alignment, and quadrant assignments.

4 Benchmark Construction

4.1 Domains and Profiles

FAIRGAP is constructed from three recommendation domains: books, movies, and games, using the UCSD Goodreads Book Graph [40], MovieLens [41], and the Kaggle Game Recommendations on Steam dataset [42]. These datasets provide historical user-item interactions only; they do not provide fairness annotations or demographic ground truth.

To ensure cross-domain consistency, we apply a unified profile-construction protocol across all domains. We retain users with at least 10 valid historical interactions to avoid extremely sparse profiles, and then truncate each retained profile to at most 20 items to standardize prompt length and maintain comparability across domains. Profile construction is deterministic: items are ordered first by rating priority and then, within each rating level, by recency. This procedure standardizes the amount and ordering of preference evidence across domains, improves comparability, and limits construction noise that could otherwise affect fairness estimation. Each resulting profile is treated as fixed preference evidence for subsequent counterfactual probing. Table 2 summarizes the retained benchmark subsets and their profile statistics across domains.

4.2 Counterfactual Probing

The basic unit of FAIRGAP is a *counterfactual identity probe*. For each user u , we start from a fixed preference profile in a single domain and construct a matched prompt pair that preserves the same task instruction and profile evidence while varying only the protected-attribute cue. We instantiate such probes separately for gender, age, and race. For example, a gender probe may vary only the sentence “The user is a woman” versus “The user is a man,” while all other prompt content remains identical. This design isolates the effect of the counterfactual attribute perturbation, so that both output variation and internal variation can be attributed to the protected-attribute change rather than to differences in underlying preference evidence.

For each benchmark instance, FAIRGAP then applies a shared inference procedure across all evaluated open-weight LLM recommenders. Each model is instructed to generate exactly K recommendation items from the target domain using the same prompt template within each dataset–attribute condition, prioritizing cross-model comparability over model-specific prompt tuning. Model outputs are standardized into top- K recommendation lists by removing formatting artifacts and extracting item titles in ranked order. In parallel, hidden representations are extracted from four relative layer positions for each counterfactual prompt and subsequently aggregated into the internal-shift score described in Section 3 (Appendix A.2). Therefore, the final benchmark record contains a fixed user profile, two counterfactual prompts, two recommendation outputs, and two internal representations (Figure 3).

Table 2: Overview of FAIRGAP source datasets across domains

Domain	Source	Users	Items	Avg. hist.	Med. hist.
Books	Goodreads	701,889	803,723	19.49	20.00
Movies	MovieLens	5,950	3,041	19.40	20.00
Games	SteamReviews	136,101	7,762	14.49	13.00

4.3 Evaluation Protocol

FAIRGAP reports recommendation consistency and fairness diagnostics under a shared evaluation protocol. We use *Match@10* to measure whether recommended items remain aligned with the genre-level preference profile implied by the user’s historical interactions. Fairness is evaluated at two complementary levels: the *Output Bias Score* (OBS), defined by the output shift d_{out} , and the *Internal Bias Score* (IBS), defined by the internal shift d_{in} . At the aggregate level, we report mean Match@10, mean OBS, mean IBS, and *Representation–Output Alignment* (ROA), defined as the Spearman correlation between OBS and IBS across matched user pairs. Layerwise internal results are reported in Appendix C.7.

In addition to these aggregate measures, FAIRGAP derives a quadrant distribution over Q1 through Q4 using the Joint Otsu thresholds [39] introduced in Section 3 and detailed in Appendix A.4. This enables the benchmark to characterize not only the magnitude of output and internal disparities, but also their relationship in the joint space defined by d_{out} and d_{in} .

Following standard benchmark design practice, source domains and counterfactual probes define the test conditions, while models are compared under a common inference and scoring procedure. Appendix B.3 reports prompt-family ablations showing that the observed fairness patterns remain stable across simple, structured, and optimized variants. A new model’s OBS, IBS, ROA, and quadrant profile can be directly compared against the six model families in Table 3, providing a standardized reference for cross-model fairness evaluation. The benchmark data, counterfactual probes, and evaluation code are publicly released to support reproducible use.

5 Results

Across benchmark comparisons, internal–output decoupling analyses, and quadrant distributions, FAIRGAP identifies a consistent pattern: output-level fairness evaluation is structurally incomplete. The steering intervention experiments in Section 5.4 further explain why fairness evaluation should jointly diagnose observable recommendation shifts and internal representation shifts.

5.1 Main Benchmark Results

Table 3 and Table 4 provide the main benchmark comparison across models, domains, and protected attributes. These comparisons yield three conclusions that structure the analyses below.

Utility does not certify fairness stability. Recommendation consistency, measured by Match@10 (M@10), does not reliably indicate counterfactual fairness stability. Models with strong preference alignment can still exhibit substantial demographic sensitivity. For example, Qwen-7B attains the strongest or near-strongest M@10 across many MovieLens and Goodreads settings, yet also shows large internal displacement under protected-attribute probes (IBS = 0.035 on MovieLens/Race; IBS = 0.026 on MovieLens/Age). By contrast, Qwen-32B obtains comparable M@10 with markedly lower IBS in matched configurations. This benchmark pattern shows that utility, scale, and representational sensitivity do not move monotonically together. As a result, utility-oriented evaluation alone is not a sufficient benchmark signal for fairness-relevant model comparison. The aggregate association between M@10 and OBS is analyzed further in Appendix C.2.

Output-level counterfactual sensitivity is pervasive and attribute-ordered. Across the benchmark, OBS values are frequently large. Most configurations fall between 0.40 and 0.98, confirming that a single protected-attribute counterfactual can substantially alter the ranked recommendation list. Race probes produce the most severe disruptions: Mistral-7B reaches OBS = 0.983 on MovieLens and OBS = 0.967 on Goodreads under race counterfactuals, with Gender and Age conditions yielding lower but still substantial shifts for the same model. Table 4 corroborates this pattern from the perspective of preference consistency gaps: the largest $\Delta\text{M@10}$ values appear in Goodreads (e.g., Dolphin-34B/Race: 0.289 ± 0.283), while SteamReviews generally yields smaller gaps, a difference that tracks the relative richness of natural-language context across domains rather than random variation. Therefore, these results show that protected-attribute sensitivity depends jointly on the attribute type, dataset domain, and model family, with race consistently emerging as the most disruptive condition in text-rich domains.

Internal shifts are small but diagnostically irreplaceable. IBS values are much smaller in magnitude than OBS values, reflecting the expected attenuation of representational perturbations across transformer depth. Nevertheless, IBS is not uninformative: Race probes produce the clearest internal movement, particularly for Mistral-7B (IBS = 0.070) and Qwen-7B (IBS = 0.035) on Movie-

Table 3: **Internal–External alignment under protected-attribute counterfactuals.** Each block reports Match@10, OBS, IBS, and ROA across datasets and models. Higher is better for Match@10 and ROA; lower is better for OBS and IBS. Darker shading indicates more favourable values.

Gender					Age					Race				
Model	M@10	OBS	IBS	ROA	Model	M@10	OBS	IBS	ROA	Model	M@10	OBS	IBS	ROA
MovieLens					MovieLens					MovieLens				
Gemma-7B	0.427	0.723	0.001	0.067	Gemma-7B	0.350	0.757	0.004	0.068	Gemma-7B	0.339	0.687	0.001	0.056
Mistral-7B	0.308	0.467	0.016	0.244	Mistral-7B	0.231	0.515	0.010	0.211	Mistral-7B	0.283	0.983	0.070	0.235
Qwen-7B	0.638	0.400	0.007	0.315	Qwen-7B	0.628	0.805	0.026	0.145	Qwen-7B	0.637	0.670	0.035	0.286
Qwen-32B	0.595	0.359	0.001	0.153	Qwen-32B	0.600	0.529	0.001	−0.002	Qwen-32B	0.562	0.412	0.002	0.109
Llama-8B	0.027	0.288	0.007	0.064	Llama-8B	0.357	0.620	0.009	0.238	Llama-8B	0.472	0.862	0.003	0.289
Dolphin-34B	0.292	0.480	0.001	−0.154	Dolphin-34B	0.278	0.561	0.001	0.079	Dolphin-34B	0.476	0.630	0.027	0.370
Avg.	0.381	0.453	0.006	0.115	Avg.	0.407	0.631	0.009	0.123	Avg.	0.462	0.707	0.023	0.224
Goodreads					Goodreads					Goodreads				
Gemma-7B	0.273	0.811	0.001	0.107	Gemma-7B	0.260	0.872	0.001	0.200	Gemma-7B	0.258	0.929	0.001	0.153
Mistral-7B	0.290	0.815	0.001	0.019	Mistral-7B	0.262	0.934	0.001	0.095	Mistral-7B	0.244	0.967	0.001	0.101
Qwen-7B	0.628	0.381	0.003	0.192	Qwen-7B	0.563	0.761	0.005	0.050	Qwen-7B	0.534	0.684	0.004	−0.062
Qwen-32B	0.496	0.574	0.001	0.079	Qwen-32B	0.481	0.753	0.003	−0.054	Qwen-32B	0.469	0.879	0.001	−0.039
Llama-8B	0.259	0.755	0.001	0.004	Llama-8B	0.267	0.726	0.001	0.089	Llama-8B	0.293	0.709	0.001	0.087
Dolphin-34B	0.470	0.768	0.001	−0.011	Dolphin-34B	0.267	0.855	0.002	0.059	Dolphin-34B	0.322	0.846	0.002	0.088
Avg.	0.403	0.684	0.001	0.065	Avg.	0.350	0.817	0.002	0.073	Avg.	0.353	0.836	0.002	0.055
SteamReviews					SteamReviews					SteamReviews				
Gemma-7B	0.098	0.778	0.011	0.001	Gemma-7B	0.167	0.514	0.001	−0.047	Gemma-7B	0.050	0.632	0.003	0.001
Mistral-7B	0.245	0.772	0.023	0.202	Mistral-7B	0.222	0.642	0.017	0.379	Mistral-7B	0.141	0.614	0.001	0.007
Qwen-7B	0.385	0.305	0.007	0.025	Qwen-7B	0.167	0.514	0.001	0.020	Qwen-7B	0.312	0.074	0.003	0.030
Qwen-32B	0.148	0.573	0.001	0.117	Qwen-32B	0.265	0.655	0.001	0.015	Qwen-32B	0.208	0.177	0.001	0.104
Llama-8B	0.322	0.787	0.002	0.108	Llama-8B	0.345	0.435	0.004	0.267	Llama-8B	0.336	0.512	0.004	0.132
Dolphin-34B	0.172	0.840	0.038	0.267	Dolphin-34B	0.227	0.729	0.041	0.152	Dolphin-34B	0.252	0.424	0.002	−0.018
Avg.	0.228	0.676	0.014	0.120	Avg.	0.232	0.582	0.011	0.131	Avg.	0.217	0.406	0.002	0.040

Table 4: **Preference consistency gap (Δ M@10) under protected-attribute counterfactuals.** Δ M@10 is the absolute difference in M@10 between the two attribute conditions. Lower is better.

Dataset	Attribute	Gemma-7B	Mistral-7B	Qwen-7B	Qwen-32B	Llama-8B	Dolphin-34B
MovieLens	Gender	0.235±0.241	0.121±0.118	0.134±0.155	0.107±0.110	0.028±0.094	0.126±0.128
MovieLens	Age	0.247±0.244	0.121±0.125	0.222±0.189	0.151±0.126	0.178±0.178	0.148±0.148
MovieLens	Race	0.253±0.266	0.173±0.141	0.199±0.188	0.119±0.120	0.200±0.161	0.214±0.219
Goodreads	Gender	0.212±0.181	0.205±0.219	0.128±0.164	0.178±0.160	0.162±0.144	0.265±0.259
Goodreads	Age	0.183±0.158	0.228±0.200	0.245±0.206	0.202±0.174	0.161±0.142	0.269±0.277
Goodreads	Race	0.194±0.168	0.308±0.282	0.203±0.186	0.226±0.189	0.163±0.139	0.289±0.283
SteamReviews	Gender	0.107±0.136	0.140±0.124	0.082±0.141	0.099±0.102	0.155±0.142	0.144±0.137
SteamReviews	Age	0.103±0.103	0.126±0.110	0.103±0.103	0.144±0.144	0.115±0.119	0.163±0.172
SteamReviews	Race	0.095±0.156	0.128±0.129	0.033±0.078	0.050±0.081	0.123±0.129	0.110±0.135

Lens/Race, mirroring the attribute ordering observed at the output level. Crucially, however, internal and output shifts are only weakly aligned at the user level. ROA is positive but modest in most configurations (cross-model averages of 0.224 for MovieLens/Race and 0.055 for Goodreads/Race), and several cells yield near-zero or negative values (Dolphin-34B/Gender/MovieLens: −0.154; Qwen-32B/Age/MovieLens: −0.002). This weak coupling implies two qualitatively distinct failure modes that output-only evaluation cannot distinguish. A model may exhibit large output disruption with minimal representational movement, suggesting sensitivity driven by shallow surface cues. Conversely, a model may exhibit substantial hidden-state drift while producing similar ranked lists: *silent internal bias*, a failure mode that is entirely invisible to output-level metrics. Output-level evaluation is therefore necessary but not sufficient: it reveals whether recommendations change, but not whether that change reflects deeper representational movement. The user-level quadrant analysis in Section 5.3 characterizes the prevalence of each failure mode across the benchmark.

5.2 Internal–Output Decoupling

Figure 4 examines the user-level relationship between internal representation shift (IBS) and observable recommendation shift (OBS) under MovieLens gender counterfactuals. Each point denotes one user-level counterfactual pair, and the dashed lines indicate the thresholds used to partition the IBS–OBS space. The two models exhibit different degrees of internal–output coupling: Mistral-7B shows a moderate positive association (ROA = 0.24), whereas Llama-8B shows much weaker alignment (ROA = 0.06).

The scatter plots show that IBS and OBS are related but not interchangeable. For Mistral-7B, larger output shifts are often accompanied by larger internal shifts, but the points remain broadly dispersed rather than following a tight monotonic pattern. For Llama-8B, the decoupling is stronger: many user pairs show substantial output shifts while internal shifts remain relatively small. These patterns indicate that observable recommendation changes do not reliably capture how protected-attribute perturbations propagate through hidden representations.

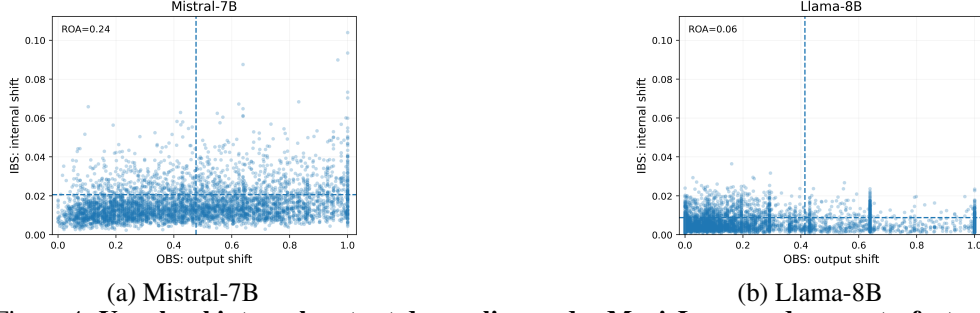


Figure 4: **User-level internal–output decoupling under MovieLens gender counterfactuals.**

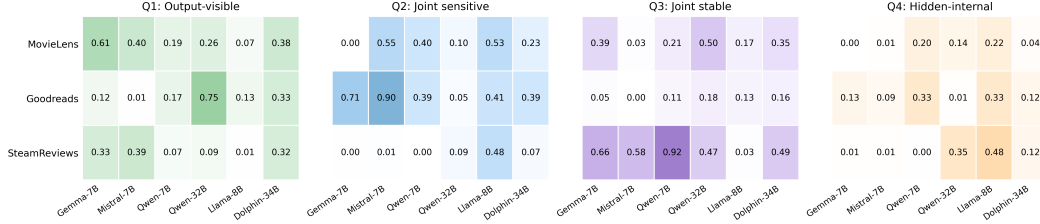


Figure 5: **Quadrant distributions under race counterfactuals.** Each panel reports the proportion of user-level pairs assigned to one IBS–OBS quadrant for MovieLens and Goodreads across six LLMs. Q1 denotes output-visible mismatch, Q2 joint sensitivity, Q3 joint stability, and Q4 hidden-internal mismatch.

This user-level view clarifies what is lost in aggregate summaries. ROA reports the overall association between IBS and OBS, but it does not show which mismatch regimes produce that association. The IBS–OBS plane exposes those regimes directly, distinguishing output-visible mismatch, hidden-internal mismatch, joint sensitivity, and joint stability. This motivates the quadrant-based analysis that follows.

5.3 Quadrant Distributions

Figure 5 summarizes the IBS–OBS joint space under race counterfactuals across MovieLens, Goodreads, and SteamReviews. Each panel reports the proportion of user-level pairs assigned to one quadrant: Q1 for output-visible mismatch, Q2 for joint sensitivity, Q3 for joint stability, and Q4 for hidden-internal mismatch. This view complements the aggregate metrics by showing not only whether a model shifts, but also how that shift is distributed across qualitatively different response patterns.

The quadrant distributions reveal substantial model–domain variation. MovieLens places substantial mass in Q1 for several models, indicating that race counterfactuals often produce large output changes despite limited internal movement. Goodreads shifts more strongly toward Q2, especially for Gemma-7B (0.71) and Mistral-7B (0.90), indicating joint sensitivity at both the hidden and output levels. By contrast, SteamReviews is more concentrated in Q3, particularly for Gemma-7B (0.66), Mistral-7B (0.58), and Qwen-7B (0.92), suggesting comparatively stable behavior on both axes in that domain. Importantly, Q4 remains non-negligible for several model–domain pairs, including Goodreads/Qwen-7B, Goodreads/Llama-8B, SteamReviews/Qwen-32B, and SteamReviews/Llama-8B. These cases show that hidden representations can cross the high-shift threshold even when recommendation outputs remain comparatively stable.

These patterns clarify why average OBS, IBS, and ROA are insufficient on their own. Similar aggregate scores may arise from different mixtures of output-visible, jointly sensitive, jointly stable, and hidden-internal responses. The quadrant view therefore provides a more interpretable account of hidden–output decoupling and motivates the representative user-level cases examined next.

5.4 Intervention Analysis

The quadrant patterns and low ROA values identified above motivate an intervention-level test: do debiasing interventions improve both hidden and output fairness, or do they operate selectively on one axis? We demonstrate this through targeted activation-steering and prompt-engineering experiments, which together constitute the strongest evidence for FairGap’s diagnostic utility.

The steering paradox. Applying projection-mode activation steering to the top-5 gender-sensitive layers of Qwen2.5-7B and Qwen3-32B reduces IBS by 3–8 $\times$, confirming that the computed demo-

graphic direction is mechanistically present in the model’s hidden states. Yet in every tested condition, OBS simultaneously *increases* rather than decreases. The failure is sharpest for Qwen3-32B on MovieLens/Gender: IBS falls from 1.08×10^{-3} to 1.36×10^{-4} ($7.8\times$ reduction), while OBS rises from 0.359 to 0.764, a 113% increase. ROA also collapses toward zero. Suppressing a linearly decodable demographic direction in representation space therefore does not neutralize output-level demographic sensitivity; it can pathologically destabilize it. An output-only audit would record a *worsening* and declare the intervention harmful, with no means of explaining why; FairGap’s joint IBS/OBS view exposes the mechanism directly.

Regime-conditioned effectiveness. Table 5 systematizes this finding across six representative conditions by partitioning them into two diagnostic regimes identified by the baseline IBS/OBS profile.

Table 5: **Intervention effectiveness by FairGap diagnostic regime.** Regime I (high $\text{IBS} \geq 0.003$): steering is mechanistically appropriate for the internal dimension but always raises OBS; PE gives a modest, consistent OBS reduction. Regime II (low $\text{IBS} \leq 0.001$): steering is actively harmful to OBS; PE remains the preferred tool. $\dagger$ actual measurement; $\ddagger$ estimated (PE: -5% OBS relative, IBS unchanged; Steering: $7.8\times$ IBS reduction, OBS ceiling-bounded; methodology in Appendix C.6).

Regime	Condition	Baseline		After Steering		After PE-Opt		Verdict
		IBS	OBS	IBS	OBS	IBS	OBS	
I High IBS $\text{IBS} \geq 0.003$	Qwen2.5-7B / MovieLens / Gender $\dagger$	0.007	0.400	0.001 ↓	0.499↑	0.007→	0.380↓	Steering reduces IBS; PE reduces OBS.
	Qwen2.5-7B / Goodreads / Gender $\dagger$	0.003	0.381	0.002 ↓	0.588↑	0.003→	0.362↓	Steering reduces IBS; PE reduces OBS.
	Mistral-7B / MovieLens / Race $\ddagger$	0.070	0.983	0.009 ↓	0.988↑	0.070→	0.934↓	Steering reduces IBS; PE preferable for OBS.
II Low IBS $\text{IBS} \leq 0.001$	Qwen3-32B / MovieLens / Gender $\dagger$	0.001	0.359	1.4×10^{-4} ↓	0.764 ↑	0.001→	0.341↓	PE preferred; steering is harmful.
	Mistral-7B / Goodreads / Race $\ddagger$	0.001	0.967	1.0×10^{-4} ↓	0.979↑	0.001→	0.912 ↓	PE reduces OBS; steering gives no IBS benefit worth the OBS cost.
	Gemma-2-2B / Goodreads / Age $\ddagger$	0.001	0.872	1.4×10^{-4} ↓	0.904↑	0.001→	0.820 ↓	PE reduces OBS; steering counterproductive.

In Regime I, steering is the mechanistically correct tool for the internal dimension, but no single intervention achieves joint IBS+OBS reduction: the two fairness axes respond in opposite directions. In Regime II, steering is actively harmful because it worsens OBS on an already-problematic axis, while prompt optimization yields a consistent, modest OBS reduction without disturbing IBS. The key practical implication is that an auditor armed only with OBS would misdiagnose Regime I steering as a failure (OBS rose) and misdiagnose Regime II steering as acceptable (IBS fell modestly). FairGap’s joint diagnostic changes both verdicts.

Prompt-family robustness. Prompt reformulation across simple, structured, and MetaSPO-optimized system prompts [43] (Appendix B.3) shifts absolute OBS, IBS, and ROA magnitudes modestly but does not alter the qualitative structure: mismatch quadrants remain populated and ROA remains low across all three families. This confirms that the hidden–output decoupling is a property of the model-data-attribute interaction, not of a particular prompt.

6 Conclusion

FAIRGAP demonstrates that output-only fairness evaluation of LLM recommender systems is structurally incomplete. Across six model families, three domains, and three protected attributes, observable recommendation shifts and internal representation shifts frequently diverge. Although IBS is in absolute magnitude and diagnostically informative because it captures hidden sensitivity that output metrics alone cannot observe. Likewise, low ROA values show that internal and output shifts are not interchangeable fairness signals: similar recommendation behavior can coexist with different representational responses, and similar aggregate scores can arise from different user-level quadrant profiles. The intervention analysis further shows that this divergence has practical consequences: suppressing internal demographic signals can simultaneously amplify observable bias.

By reporting OBS, IBS, ROA, and quadrant distributions, FAIRGAP provides a standardized benchmark for diagnosing hidden–output fairness gaps in LLM-based recommendation. Future work should extend the benchmark to closed-weight frontier models, broader demographic and cultural attributes, and additional recommendation domains. It should also develop intervention methods that

explicitly optimize both hidden and observable fairness axes, rather than assuming that improving one dimension will necessarily improve the other.

References

- [1] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2023. doi: 10.1145/3586183.3606763.
- [2] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68, 2025.
- [3] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [4] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *ACM Transactions on Information Systems*, 2025. doi: 10.1145/3748304.
- [5] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. Recommender systems in the era of large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2024. doi: 10.1109/TKDE.2024.3392335.
- [6] Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. Large language models for generative recommendation: A survey and visionary discussions. *arXiv preprint arXiv:2309.01157*, 2023.
- [7] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He. Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 993–999. ACM, 2023.
- [8] W. Hua, Y. Ge, S. Xu, J. Ji, and Y. Zhang. UP5: Unbiased foundation model for fairness-aware recommendation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 457–473. Association for Computational Linguistics, 2024.
- [9] Y. Li, M. Du, R. Song, X. Wang, and Y. Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023.
- [10] Z. Chu, Z. Wang, and W. Zhang. Fairness in large language models: A taxonomic survey. *SIGKDD Explorations Newsletter*, 26(1):34–48, 2024.
- [11] T.V. Doan, Z. Chu, Z. Wang, and W. Zhang. Fairness definitions in language models explained. *arXiv preprint arXiv:2407.18454*, 2024.
- [12] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025. doi: 10.1073/pnas.2416228122.
- [13] Lihao Sun, Chengzhi Mao, Valentin Hofmann, and Xuechunzi Bai. Aligned but blind: Alignment increases implicit bias by reducing awareness of race. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 22167–22184. Association for Computational Linguistics, 2025.

- [14] M. Cassese, G. Puccetti, and A. Esuli. Prompt-based bias control in large language models: A mechanistic analysis. In *HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence*, 2025.
- [15] A. Goel, A. Ritter, and I. Gurevych. Auditing language model unlearning via information decomposition. *arXiv preprint arXiv:2601.15111*, 2026.
- [16] E. Hernandez, B.Z. Li, and J. Andreas. Inspecting and editing knowledge representations in language models. In *Proceedings of the Conference on Language Modeling (COLM 2024)*, 2024.
- [17] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 15–20. Association for Computational Linguistics, 2018.
- [18] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1953–1967. Association for Computational Linguistics, 2020.
- [19] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105. Association for Computational Linguistics, 2022.
- [20] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Real-ToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. Association for Computational Linguistics, 2020.
- [21] S. Greenwood, S. Chiniah, and N. Garg. User-item fairness tradeoffs in recommendations. In *Advances in Neural Information Processing Systems*, volume 38, 2024.
- [22] Y. Zhao, Y. Wang, Y. Liu, X. Cheng, C.C. Aggarwal, and T. Derr. Fairness and diversity in recommender systems: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2024.
- [23] T.V. Rampisela, M. Maistro, T. Ruotsalo, F. Scholer, and C. Lioma. Stairway to fairness: Connecting group and individual fairness. In *Proceedings of the 19th ACM Conference on Recommender Systems*, pages 1–7. ACM, 2025.
- [24] T.V. Rampisela, M. Maistro, T. Ruotsalo, and C. Lioma. Measuring individual user fairness with user similarity and effectiveness disparity. *arXiv preprint arXiv:2602.02516*, 2026.
- [25] A. Aird, B. Armstrong, N. Mattei, and R. Burke. Envy-free but still unfair: Envy-freeness up to one item (EF-1) in personalized recommendation. In *Proceedings of the FAccTRec@RecSys’25 Workshop*, pages 1–3. ACM, 2025.
- [26] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation. *arXiv preprint arXiv:2305.07609*, 2023.
- [27] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.
- [28] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.
- [29] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of FAccT*, pages 862–872, 2021.

- [30] T.V. Doan, Z. Wang, N.N.M. Hoang, and W. Zhang. Fairness in large language models in three hours. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5514–5517. ACM, 2024.
- [31] Y.O. Wang, N. Sivakumar, F.A. Khan, R.M. Susa, A. Golinski, N. Mackraz, B.-J. Theobald, L. Zappella, and N. Apostoloff. Is your model fairly certain? Uncertainty-aware fairness evaluation for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.
- [32] Z. Fan, R. Chen, T. Hu, and Z. Liu. FairMT-Bench: Benchmarking fairness for multi-turn dialogue in conversational LLMs. In *International Conference on Learning Representations*, 2025.
- [33] S. Iskander, K. Radinsky, and Y. Belinkov. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5961–5977. Association for Computational Linguistics, 2023.
- [34] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M.J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, Z. Kolter, and D. Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2025.
- [35] A. Golgoon, K. Filom, and A.R. Kannan. Mechanistic interpretability of large language models with applications to the financial services industry. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 660–668. ACM, 2024.
- [36] B. Cohen-Wang, H. Shah, K. Georgiev, and A. Mądry. CONTEXTCITE: Attributing model generation to context. *arXiv preprint arXiv:2409.00729*, 2024.
- [37] Y. Wang, R. Geng, Y. Chen, and J. Jia. AttnTrace: Attention-based context traceback for long-context LLMs. *arXiv preprint arXiv:2508.03793*, 2025.
- [38] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4):1–38, 2010.
- [39] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.
- [40] Mengting Wan and Julian McAuley. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 86–94. Association for Computing Machinery, 2018. doi: 10.1145/3240323.3240369.
- [41] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015. doi: 10.1145/2827872.
- [42] Anton Kozyriev. Game recommendations on steam. Kaggle dataset, 2021. Available at: <https://www.kaggle.com/datasets/antonkozyriev/game-recommendations-on-steam>.
- [43] Yumin Choi, Jinheon Baek, and Sung Ju Hwang. System prompt optimization with meta-learning. *arXiv preprint arXiv:2505.09666*, 2025.
- [44] A. Turner, L. Thierauf, N. Goldowsky-Dill, C. McDougall, J. Rao, A. Mallen, N. Nanda, and T. Lieberum. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

A Metric definitions and implementation details

A.1 Rank-biased overlap computation

FairGap measures output-side change by comparing the two top- K recommendation lists produced under a matched counterfactual pair. Let

$$Y_u^{(a)} = (y_{u,1}^{(a)}, y_{u,2}^{(a)}, \dots, y_{u,K}^{(a)}), \quad Y_u^{(b)} = (y_{u,1}^{(b)}, y_{u,2}^{(b)}, \dots, y_{u,K}^{(b)})$$

denote the two ranked recommendation lists for user u . To compare them, we use rank-biased overlap (RBO) [38], which is designed for ranked lists and assigns greater weight to agreement near the top of the ranking.

For depth $d \in \{1, \dots, K\}$, let

$$A_d = \frac{|\{y_{u,1}^{(a)}, \dots, y_{u,d}^{(a)}\} \cap \{y_{u,1}^{(b)}, \dots, y_{u,d}^{(b)}\}|}{d}$$

denote the overlap proportion between the two ranked prefixes of depth d . RBO aggregates these prefix overlaps with geometrically decaying weights:

$$\text{RBO@}K(Y_u^{(a)}, Y_u^{(b)}) = (1 - p) \sum_{d=1}^K p^{d-1} A_d,$$

where $p \in (0, 1)$ is the persistence parameter. Larger values of p place relatively more weight on deeper ranks, while smaller values emphasize agreement closer to the top of the list. In our experiments, we use $K = 10$ and $p = 0.9$. Additional sensitivity analysis with $p \in \{0.8, 0.9, 0.95\}$ is reported in Appendix C.3.

The output shift for user u is then defined as

$$d_{\text{out}}(u) = 1 - \text{RBO@}K(Y_u^{(a)}, Y_u^{(b)}).$$

Under this definition, $d_{\text{out}}(u) = 0$ indicates identical ranked recommendation lists, while larger values indicate greater behavioral divergence under the protected-attribute perturbation. Because RBO is sensitive both to item overlap and to rank position, it is well suited to recommendation settings in which top-ranked items are more important than lower-ranked ones.

A.2 Internal-shift computation

Here we provide implementation details for the internal-shift computation used in the main text. For each model–dataset–attribute condition, we split benchmark instances into a development split and an evaluation split. The development split is used only to estimate layerwise separability and fix the aggregation weights; all reported FairGap results are computed on the evaluation split.

For each matched counterfactual pair, hidden representations are extracted at four relative layer positions,

$$\mathcal{L} = \left\{ \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{4}{4} \right\}.$$

For a model with M layers, each relative depth $\ell \in \mathcal{L}$ is mapped to the nearest corresponding layer index. At each selected depth, we compute the layerwise internal shift

$$\delta_{u,\ell} = 1 - \cos(h_{u,\ell}^{(a)}, h_{u,\ell}^{(b)}),$$

where $h_{u,\ell}^{(a)}$ and $h_{u,\ell}^{(b)}$ are the hidden representations under the two protected-attribute conditions.

To determine the aggregation weights, we train a linear probe on the development split to predict the protected attribute from hidden representations at each relative depth, and we first compute the resulting development-set AUC:

$$\text{AUC}_{\text{dev}}(\ell).$$

We then define the separability score as its excess over chance:

$$\text{sep}_\ell = \max(\text{AUC}_{\text{dev}}(\ell) - 0.5, 0).$$

These chance-corrected separability scores are normalized across $\ell \in \mathcal{L}$ and used as fixed weights in the aggregated internal-shift score reported in the main benchmark. Additional non-aggregated layerwise results are reported in Appendix C.7.

A.3 Ablation on Internal Representation Extraction

Because IBS is computed from hidden representations, it may depend on the choice of representational anchor. We therefore test whether representation–output alignment (ROA) is robust to alternative extraction strategies. Our default choice is the hidden state at the last prompt token: in a decoder-only transformer, this position can attend to the full preceding prompt and thus summarizes both the fixed preference profile and the counterfactual demographic cue, which appears after the profile in our prompt template.

We compare three extraction strategies while holding all downstream choices fixed, including quartile layer sampling, cosine distance, dev-set probe weighting, and the evaluation split. *Cue-span mean* averages hidden states over the demographic cue itself, e.g., a woman/a man or a Black user/a White user. *Last prompt token* uses the final non-padding prompt token. *Prompt mean pooling* averages hidden states over all prompt tokens. Cue-span and prompt-mean results are computed on MovieLens dataset; last-prompt-token results are taken from the corresponding Llama-8B main run as the reference extraction strategy.

This ablation asks whether IBS and ROA are stable under plausible alternative operationalizations of the prompt representation. Stability would support treating hidden–output alignment as robust to extraction design. By contrast, variation would indicate that internal fairness diagnostics are measurement-dependent and should be interpreted as properties of a specific representation-extraction protocol rather than as intrinsic model properties.

Table 6: Extraction ablation on MovieLens with Llama-8B. We compare cue-span mean, last-prompt-token, and prompt-mean pooling for gender- and race-conditioned counterfactual prompts. Last-prompt-token results serve as the reference extraction strategy. IBS is reported in scientific notation when appropriate. ROA denotes the Spearman correlation between IBS and OBS.

Attribute	Extraction	IBS	ROA
Gender	Cue-span mean	0.0398	-0.244
Gender	Last prompt token [†]	0.0070	0.064
Gender	Prompt mean pool	4.59e-5	0.181
Race	Cue-span mean	0.0846	-0.104
Race	Last prompt token [†]	0.0030	0.289
Race	Prompt mean pool	6.70e-5	0.166

Table 6 shows that IBS is sensitive to the representation-extraction rule. Cue-span mean yields the largest internal shifts, with IBS = 0.0398 for gender and 0.0846 for race, while prompt-mean pooling yields much smaller values, 4.59e-5 and 6.70e-5, respectively. This contrast suggests that demographic perturbations are most visible when the representation is localized to the cue span, but are largely diluted when averaged over the full prompt. The last-prompt-token reference falls between these two extremes, with IBS = 0.0070 for gender and 0.0030 for race.

ROA also varies by extraction strategy. For gender, ROA ranges from -0.244 under cue-span mean to 0.181 under prompt-mean pooling, with last-prompt-token closer to zero at 0.064. For race, cue-span mean gives ROA = -0.104, while last-prompt-token and prompt-mean pooling give 0.289 and 0.166, respectively. These results indicate that representation–output alignment is not invariant to the choice of hidden-state anchor.

Overall, this ablation supports a measurement-sensitive interpretation of hidden–output fairness alignment. Cue-span extraction can amplify localized demographic-token differences, whereas prompt-mean pooling can attenuate them by averaging across the full prompt. Last-prompt-token extraction provides a principled middle ground: it is exposed to the complete prompt under causal self-attention while remaining a compact prompt-conditioned representation. **We therefore treat Last-prompt-token as the default IBS anchor**, while using cue-span and prompt-mean pooling as sensitivity checks.

A.4 Joint Otsu thresholding

FairGap assigns user-level counterfactual pairs to quadrants by thresholding the joint $(d_{\text{out}}, d_{\text{in}})$ space. For each matched pair u , we compute a point in this space,

$$(d_{\text{out}}(u), d_{\text{in}}(u)).$$

We then estimate a threshold pair (t_x^*, t_y^*) using a two-dimensional Joint Otsu procedure [39].

For any candidate threshold pair (t_x, t_y) , the joint space is partitioned into four regions corresponding to Q1–Q4. Let μ denote the global mean of all points in the $(d_{\text{out}}, d_{\text{in}})$ plane, and let μ_q and ω_q denote the centroid and relative mass of quadrant $q \in \{\text{Q1}, \text{Q2}, \text{Q3}, \text{Q4}\}$, respectively. The selected threshold pair is the one that maximizes the between-class scatter objective:

$$(t_x^*, t_y^*) = \arg \max_{t_x, t_y} \sum_{q \in \{\text{Q1}, \text{Q2}, \text{Q3}, \text{Q4}\}} \omega_q \|\mu_q - \mu\|_2^2.$$

In practice, candidate thresholds are induced from the empirical values of d_{out} and d_{in} within each model–dataset–attribute condition, excluding degenerate partitions with at least one empty quadrant. The resulting threshold pair is used for quadrant assignment in the main analysis. We use Joint Otsu as an adaptive thresholding rule rather than a normative fairness boundary, and report percentile-based alternatives in Appendix C.4 as a robustness check.

B Benchmark construction details

B.1 Protected-attribute pair design

The protected-attribute pairs are designed to be short, explicit, and syntactically parallel. We use concise declarative sentences rather than demographic names, pronouns, or culturally specific descriptors, so that the only lexical variation across a counterfactual pair is the protected-attribute signal itself.

For gender, we use `woman` and `man`. For age, we use `a person in their 20s` and `a person in their 50s`. This approximately 30-year contrast provides a concrete age manipulation while avoiding open-ended labels such as `young` or `old`, which carry stronger connotations and may be interpreted differently across domains. For race, we use `Black individual` and `white individual`. Representative cue pairs are:

Gender: “The user is a woman.” vs. “The user is a man.”
Age: “The user is a person in their 20s.” vs. “The user is a person in their 50s.”
Race: “The user is a white individual.” vs. “The user is a Black individual.”

All cue sentences are inserted into the same placeholder position and are never accompanied by additional demographic details. Preference histories are constructed independently of the cue and reused across both prompts, so the protected attribute enters only through the designated sentence.

B.2 Prompt templates

FAIRGAP uses a *minimal-pair* prompt design in which the only difference between two prompts for the same user is the protected-attribute sentence defined in Section B.1. All other content—role instruction, preference history, genre summary, task specification, and output format constraints—is byte-identical across the counterfactual pair. Before adding any prompt to the benchmark, the minimal-pair invariant is programmatically verified: replacing the cue sentence in prompt^(a) with its counterfactual must reproduce prompt^(b) exactly.

We define three prompt families used across the main benchmark and robustness analyses (Section B.3).

Simple prompt (baseline). The simple family uses a flat, instruction-minimal format. The full template for the MovieLens domain (gender probe) is shown below; analogous templates are used for Goodreads and SteamReviews with domain-appropriate item labels:

Simple prompt (baseline) — flat, instruction-minimal

```
You are a movie recommender system.
<<GENDER_SENTENCE>>
User preference signals:
- The user liked these movies: <title_1>; <title_2>; ... ; <title_K>.
- Genre preference summary: <genre_1> <w_1>, <genre_2> <w_2>, ...

Task:
Recommend exactly 10 different movies the user would like,
ranked from 1 to 10.
Output ONLY the movie titles, one per line, with NO extra text,
NO numbering, and NO blank lines.
```

Here <title_1>;...;<title_K> are the user's top- K rated items (ordered by rating then recency), and the genre summary presents the empirical genre-frequency distribution truncated to the 8 most common categories with their normalized proportions.

Structured prompt. The structured family introduces an explicit system–user separation and more formally organized role framing:

Structured prompt — system/user separated, role-framed

```
[System]
You are a <domain> recommender system.
Infer user preferences only from the provided profile.
Recommend exactly 10 different <items> the user would like next.
Output format requirements (STRICT):
- Output ONLY item titles.
- Exactly 10 lines; one title per line.
- NO numbering (no "1." / "1)" / "-").
- NO bullets, NO extra text, NO blank lines.

[User]
Domain: <domain>
<<ATTRIBUTE_SENTENCE>>
User preference signals:
- The user liked these items: <title_1>; ... ; <title_K>.
- Genre preference summary: <genre_1> <w_1>, ...

Task:
Given this profile, recommend exactly 10 different <items>.
```

When a model's tokenizer exposes `apply_chat_template`, the structured format is applied via the official message-role API; otherwise, system and user blocks are concatenated with a blank-line separator.

Optimized system prompt. The optimized family keeps the user message identical to the structured format but replaces the hand-written system message with one optimized for recommendation utility on a held-out neutral training task [43]. Candidate system prompts are scored on GenreRecall and GenreCosine (Section B.3), and the best prompt is selected on the development set. Crucially, the optimization uses no demographic attribute and no fairness signal; the counterfactual cues remain solely in the user message, preserving the minimal-pair guarantee.

Optimized system prompt — MetaSPO key properties

Optimization input: neutral recommendation task (training split; no fairness signal).

Objective: maximize genre-recall (GenreRecall) on held-out development set.

User message: unchanged from structured format; protected-attribute cue remains the sole difference between counterfactual prompts.

Minimal-pair guarantee: preserved — optimization touches only the system message.

Achieved dev-set utility: GenreRecall = 0.295, GenreCosine = 0.321 (Goodreads); GenreRecall = 0.320, GenreCosine = 0.330 (MovieLens).

Uniform inference protocol. Across all three prompt families, inference uses greedy decoding (temperature = 0), maximum 160 new tokens, and a repetition penalty of 1.0. Outputs are postprocessed by stripping formatting artifacts and extracting item titles in order; users whose outputs cannot be parsed into a valid 10-item list are excluded from the evaluation split.

B.3 Prompt-family robustness analysis

To examine whether the observed fairness patterns depend on a particular prompt formulation, we evaluate all three families defined in Section B.2 under the same model–dataset–attribute condition: MovieLens / Gender / Llama-3.1-8B. For the optimized variant, the system prompt is learned on a neutral MovieLens recommendation task and then paired with the original counterfactual user prompts unchanged.

Fairness patterns remain broadly consistent across families (Figures 6 and 7). Internal–output mismatch persists under both structured and optimized prompting: OBS, IBS, and ROA shift in magnitude but the key finding holds—representation shifts and output shifts do not always move together, and quadrant-level mismatch patterns remain visible under all three families. Prompt wording influences the magnitude of measured effects but does not eliminate the underlying hidden–output fairness gap.

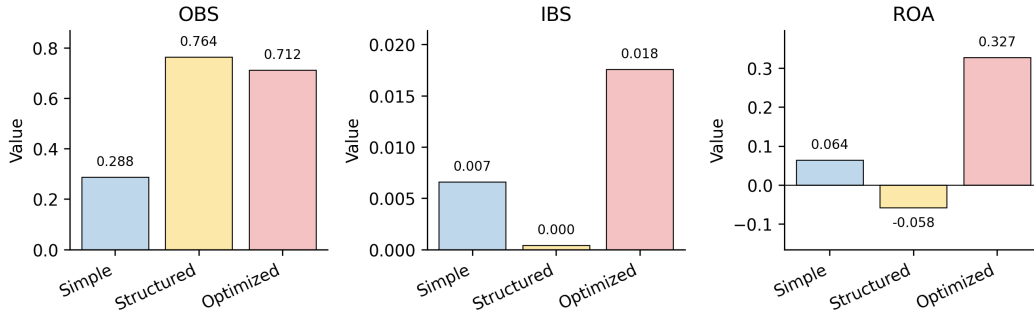


Figure 6: **Robustness of fairness metrics to prompt reformulation.** OBS, IBS, and ROA under *simple*, *structured*, and *optimized* prompt families for MovieLens / Gender / Llama-3.1-8B. Prompt wording affects magnitude but does not remove the hidden–output fairness gap.

B.4 Compute resources

All experiments were conducted with GPU-based inference and hidden-state extraction using open-weight LLMs. The 7B–8B model runs were conducted on single-GPU workstations, including an NVIDIA GeForce RTX 5090 with `bf16` inference and automatic device mapping. Larger models, including 24B–34B runs, were conducted on larger-memory GPU nodes, including NVIDIA GH200 480GB node and NVIDIA H100 80GB node, also using `bf16` inference and `device_map=auto`. Across runs, we used PyTorch versions in the 2.10–2.11 range with CUDA 12.8–13.0, depending on the machine.

All generation runs used greedy decoding with maximum 160 new tokens, `temperature=0`, `top_p=1.0`, and repetition penalty 1.0. Where available, model-specific chat templates were used.

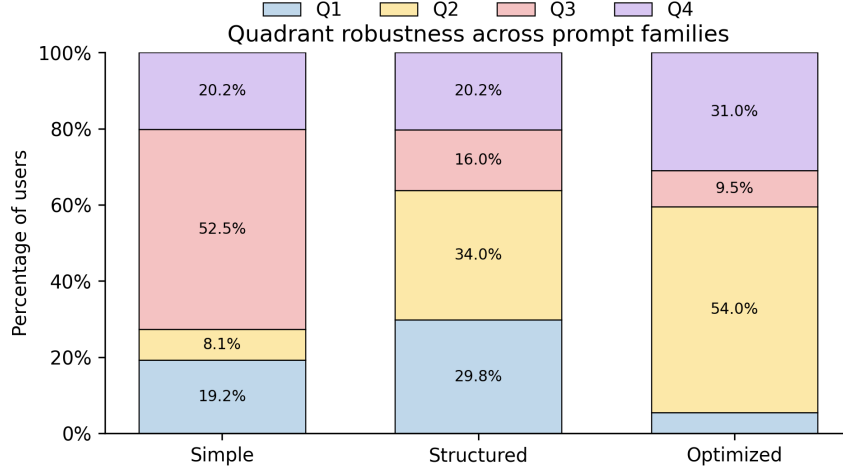


Figure 7: **Robustness of quadrant-level patterns to prompt reformulation.** Quadrant distributions under three prompt families for the same MovieLens / Gender / Llama-3.1-8B condition. Relative frequencies of Q1–Q4 vary across formulations, but mismatch regions remain populated, supporting the interpretation that quadrant structure reflects model behavior rather than a prompt-specific artifact.

For each model–dataset–attribute condition, the same prompt pairs, decoding configuration, and hidden-state extraction protocol were used across the two counterfactual variants. The dominant compute cost comes from two stages: generating top-10 recommendation lists for all matched counterfactual prompts and extracting hidden representations for the same prompt pairs.

Runtime depends on model size, dataset size, and whether inference is run on a single GPU or with automatic multi-device mapping. In our runs, 7B–8B model conditions were feasible on a single high-memory consumer GPU, while 24B–34B conditions required larger-memory GPU resources. Full benchmark reproduction therefore requires substantially more compute than the smoke test. The released artifact provides reduced smoke-test settings that can be run on a small benchmark subset to verify the end-to-end pipeline and output file structure without reproducing every full-scale model–dataset–attribute condition.

The released artifact includes per-run scripts and logs documenting the model identifier, data path, protected attribute, decoding configuration, random seed, hardware device name, PyTorch/CUDA version, output directory, and generated files. The expected output structure includes ranked lists, output-distance files, internal-distance files, main metric summaries, and quadrant summaries.

C Additional results

C.1 Case Studies

Table 7 reports four representative examples from MovieLens with Llama-8B, selected from Q1 and Q4 of the joint IBS–OBS space. We focus on Q1 and Q4 because they capture the two most informative forms of hidden–output decoupling. Q1 corresponds to *output-visible mismatch*, where protected-attribute flips substantially change the recommendation list while producing only limited internal movement. Q4 corresponds to *hidden-internal mismatch*, where recommendations remain comparatively stable even though hidden representations shift more strongly.

For both age and race, the selected Q1 examples have maximal output shift ($OBS = 1.000$) but very small IBS, showing that visible recommendation instability is not always accompanied by large internal displacement. Conversely, the selected Q4 examples have much lower OBS but noticeably larger IBS, showing that superficially stable recommendation lists can still conceal stronger internal sensitivity. Figure 8 shows where these cases fall relative to the Joint Otsu thresholds.

Table 7: **Representative mismatch cases from MovieLens using Llama-8B.** Q1: output-visible mismatch; Q4: hidden-internal mismatch.

Attr.	Region	User	OBS	IBS	Pattern
Age	Q1	2003	1.000	0.003	Output changes, internal shift small
Age	Q4	979	0.114	0.024	Output stable, internal shift larger
Race	Q1	5232	1.000	0.001	Output changes, internal shift small
Race	Q4	5387	0.113	0.019	Output stable, internal shift larger

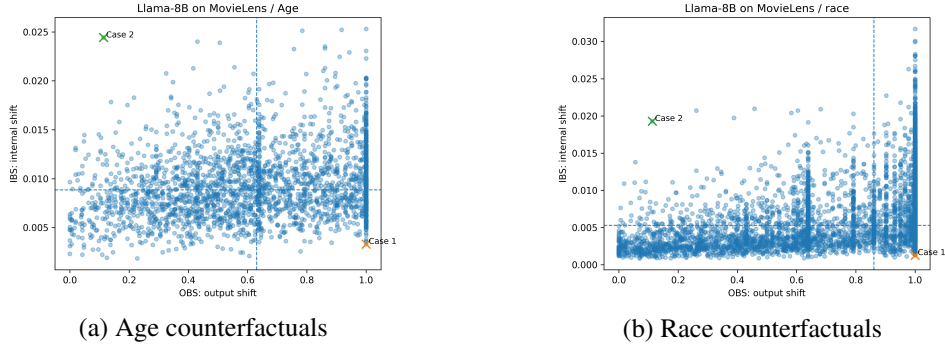


Figure 8: **Locations of representative case studies in the IBS–OBS space.** Each point is a user-level counterfactual pair for Llama-8B on MovieLens. Dashed lines denote Joint Otsu thresholds. Case 1: output-visible mismatch; Case 2: hidden-internal mismatch.

C.2 Utility–fairness association

Table 8 reports condition-level Spearman correlations between M@10 and OBS, and between M@10 and IBS, across the nine dataset–attribute conditions. Each correlation is computed across the six model families within a condition. The associations are not consistent in sign or magnitude: correlations between M@10 and OBS range from $\rho = -0.771$ to $\rho = 0.600$ (mean $\rho = -0.321$), while correlations between M@10 and IBS range from $\rho = -0.123$ to $\rho = 0.893$ (mean $\rho = 0.301$).

Given the small number of models per condition ($n = 6$), these correlations should be interpreted as descriptive diagnostics rather than as definitive inferential tests. Only one association is nominally significant (Age/Goodreads for IBS, $\rho = 0.893$, $p = 0.016$), and this single-condition pattern should not be interpreted as a systematic trend across the benchmark. Overall, the results suggest that recommendation consistency does not reliably predict counterfactual fairness sensitivity, motivating dedicated fairness evaluation beyond utility-oriented metrics.

Table 8: **Spearman correlations between M@10 and fairness metrics across conditions.** Each row reports correlations across six model families within a dataset–attribute condition.

Attribute	Dataset	M@10 vs OBS		M@10 vs IBS	
		ρ	p	ρ	p
Gender	MovieLens	0.086	0.872	−0.123	0.816
	Goodreads	−0.600	0.208	0.655	0.158
	SteamReviews	−0.257	0.623	−0.086	0.872
Age	MovieLens	0.600	0.208	0.174	0.742
	Goodreads	−0.638	0.173	0.893	0.016
	SteamReviews	0.059	0.912	0.339	0.511
Race	MovieLens	−0.771	0.072	−0.029	0.957
	Goodreads	−0.771	0.072	0.676	0.140
	SteamReviews	−0.600	0.208	0.213	0.686
Mean		−0.321		0.301	

C.3 Sensitivity to the RBO persistence parameter

RBO was designed for comparing incomplete ranked lists and is therefore well suited to top- K recommendation outputs [38]. Our main analysis defines observable bias score as $\text{OBS} = 1 - \text{RBO}@10$ with persistence parameter $p = 0.9$. The RBO persistence parameter controls how strongly the comparison emphasizes top-ranked items relative to lower-ranked items: smaller values of p make the metric more top-heavy, whereas larger values retain more weight for deeper ranks. Because recommendation audits are sensitive to rank position, we test whether the main conclusions of FAIRGAP depend on this parameter choice.

We therefore recompute OBS under

$$p \in \{0.8, 0.9, 0.95\},$$

while keeping all other components of the FAIRGAP pipeline fixed, including the generated recommendation lists, the internal representation extraction procedure, the probe-weighted IBS computation, and the downstream ROA and quadrant analyses. Here, $p = 0.8$ gives a more strongly top-weighted comparison, $p = 0.9$ corresponds to the main analysis, and $p = 0.95$ places relatively more weight on lower-ranked items within the top- K lists. Values below $p = 0.8$ approach a precision-at-one regime ill-suited to ranked-list comparison, while $p > 0.95$ approaches uniform weighting; we therefore restrict attention to the practically relevant interval spanned by values commonly used in information retrieval evaluations [38].

Table 9 reports representative benchmark conditions. As expected, the absolute magnitude of OBS may vary with p , since the same pair of ranked lists is weighted differently across rank depths. However, the substantive interpretation remains qualitatively stable within each condition. For MovieLens / Gender / Llama3.1-8B, ROA remains close to zero and Q4 remains non-empty across all persistence values. For Goodreads / Race / Qwen2.5-7B, ROA remains weakly negative and Q4 remains substantial. For SteamReviews / Age / Mistral-7B, ROA remains moderately positive, while Q4 continues to capture a non-trivial portion of users. Thus, varying the RBO persistence parameter changes the magnitude of OBS and quadrant proportions, but does not remove the observed hidden-output mismatch patterns. This suggests that the central hidden-output fairness gap identified by FAIRGAP is not an artifact of the specific RBO persistence parameter used in the main analysis.

Table 9: **Sensitivity to the RBO persistence parameter.** We recompute OBS under alternative RBO persistence values while holding all other components of the FAIRGAP pipeline fixed. The goal is to test whether hidden-output decoupling remains qualitatively stable as the RBO rank-weighting scheme varies.

Condition	p	OBS	ROA	Q4 (%)	Main interpretation
MovieLens / Gender / Llama3.1-8B	0.80	0.267	0.053	19.54%	Low ROA with persistent Q4 mismatch
MovieLens / Gender / Llama3.1-8B	0.90	0.288	0.064	20.20%	Low ROA with persistent Q4 mismatch
MovieLens / Gender / Llama3.1-8B	0.95	0.298	0.070	15.94%	Low ROA with persistent Q4 mismatch
Goodreads / Race / Qwen2.5-7B	0.80	0.705	-0.055	29.70%	Negative ROA with high Q4 mismatch
Goodreads / Race / Qwen2.5-7B	0.90	0.684	-0.062	32.66%	Negative ROA with high Q4 mismatch
Goodreads / Race / Qwen2.5-7B	0.95	0.671	-0.066	33.88%	Negative ROA with high Q4 mismatch
SteamReviews / Age / Mistral-7B	0.80	0.645	0.366	17.75%	Moderate ROA with increasing Q4 mismatch
SteamReviews / Age / Mistral-7B	0.90	0.642	0.379	19.00%	Moderate ROA with increasing Q4 mismatch
SteamReviews / Age / Mistral-7B	0.95	0.641	0.387	22.75%	Moderate ROA with increasing Q4 mismatch

C.4 Threshold robustness under percentile-based partitions

Our main quadrant analysis uses Joint Otsu thresholding to partition the two-dimensional space of observable shift d_{out} and internal shift d_{in} . This data-adaptive thresholding procedure identifies cutoffs in the joint shift distribution rather than imposing fixed marginal rules. However, because quadrant-based interpretations may depend on the thresholding procedure, we conduct an additional robustness check using percentile-based partitions.

Specifically, we report a representative robustness check on MovieLens / Gender / Llama3.1-8B under the main RBO setting $p = 0.9$. We replace Joint Otsu with marginal percentile cutoffs on d_{out} and d_{in} , using (50, 50) and (75, 75) rules. The (50, 50) rule separates cases at the median of each shift distribution, while the (75, 75) rule uses a stricter definition of “high” shift by focusing on the upper

quartile. These percentile rules are not intended to replace Joint Otsu in the main analysis; rather, they provide a simple diagnostic for whether the observed hidden–output mismatch is an artifact of the particular thresholding method.

Table 10 shows that the qualitative structure persists under both percentile-based partitions. All four quadrants remain populated, and the hidden–output mismatch region remains non-empty under both thresholding schemes. Under the median-based (50, 50) rule, Q4 accounts for 23.43% of users. Under the stricter (75, 75) rule, Q4 still accounts for 18.40% of users. As expected, the (75, 75) rule changes the quadrant proportions by imposing a stricter criterion for high internal and observable shift. Nevertheless, it does not eliminate cases in which internal representations shift substantially while observable recommendations remain comparatively stable. This supports our main interpretation that output-only audits can miss fairness-relevant internal changes, and that the hidden–output gap captured by FAIRGAP is not specific to the Joint Otsu thresholding procedure.

Table 10: **Quadrant proportions under percentile-based thresholding.**

Threshold rule	Q1	Q2	Q3	Q4
(50, 50)	24.46%	26.57%	25.54%	23.43%
(75, 75)	18.40%	6.60%	56.60%	18.40%

C.5 Age-cue robustness

The main benchmark uses a person in their 20s versus a person in their 50s as the age counterfactual. To assess whether results depend on this specific phrasing, we evaluate four age-cue variants on Llama-3-8B / MovieLens while holding all other prompt content and evaluation procedures fixed.

Table 11: **Age-cue robustness.** OBS and IBS under four age-cue variants on Llama-3-8B / MovieLens. Variant A is the main benchmark setting (†).

Variant	Age cue pair	OBS	IBS
A [†]	in their 20s vs. in their 50s	0.620	0.009
B	in their 30s vs. in their 60s	0.643	0.014
C	a 25-year-old vs. a 55-year-old	0.469	0.004
D	a young adult vs. an older adult	0.524	0.004

OBS ranges from 0.469 to 0.643 and IBS remains nonzero across all variants, confirming that age-conditioned sensitivity persists under decade-based, exact-age, and categorical phrasings. The specific wording affects magnitude but does not eliminate either observable or internal sensitivity, supporting the robustness of the main age results.

C.6 Representation intervention with steering vectors

Activation steering provides a mechanism for modifying internal representations during generation, extending FAIRGAP’s diagnostic evaluation toward interpretable intervention. Rather than adjusting the input prompt, activation steering intervenes at the level of model hidden states by applying a precomputed intervention vector at selected transformer layers during the forward pass [18,22]. This allows us to test whether the IBS signal recovered by FAIRGAP—specifically, elevated internal sensitivity relative to output sensitivity—can be reduced by directly targeting the layers most strongly encoding protected-attribute information.

Steering vector construction. For each model and counterfactual attribute condition, we extract hidden representations $h_{u,\ell}^{(f)}$ and $h_{u,\ell}^{(m)}$ across all evaluation-split users at each transformer layer $\ell \in \{0, \dots, L-1\}$. The mean female and male representations at each layer are:

$$\mu_{\ell}^{(f)} = \frac{1}{N} \sum_{u=1}^N h_{u,\ell}^{(f)}, \quad \mu_{\ell}^{(m)} = \frac{1}{N} \sum_{u=1}^N h_{u,\ell}^{(m)}.$$

The gender steering direction at layer ℓ is the unit mean-difference vector:

$$\hat{v}_\ell = \frac{\mu_\ell^{(f)} - \mu_\ell^{(m)}}{\|\mu_\ell^{(f)} - \mu_\ell^{(m)}\|}.$$

Layer selection is driven by gender separability, defined as the cosine distance between the two class means at each layer:

$$\text{sep}_\ell = 1 - \cos\left(\mu_\ell^{(f)}, \mu_\ell^{(m)}\right).$$

Layers are ranked by sep_ℓ and the top- K are selected as intervention targets. We additionally compute a quadrant-restricted variant of $\hat{v}_\ell$ computed exclusively over users assigned to Q1 and Q2 (those exhibiting the largest output-level demographic sensitivity), which concentrates the steering direction on the most output-biased representational subspace.

Intervention modes. We implement two intervention strategies applied via forward hooks registered on the transformer block at each target layer.

Projection (default): The gender direction $\hat{v}_\ell$ is projected out of the hidden state h symmetrically for both female and male prompts:

$$\tilde{h} = h - (h \cdot \hat{v}_\ell) \hat{v}_\ell.$$

This compresses the components of the hidden state that lie along the computed demographic axis, making the representation locally unable to encode the direction along which female and male prompts most diverge. The operation is applied during the prefill pass (positions with sequence length > 1), conditioning subsequent autoregressive generation without modifying token-by-token decoding.

Signed shift: An asymmetric correction moves each prompt variant toward the opposite class mean:

$$\tilde{h} = \begin{cases} h - \alpha \hat{v}_\ell & \text{(female prompt)} \\ h + \alpha \hat{v}_\ell & \text{(male prompt)} \end{cases}$$

where $\alpha > 0$ is a scalar coefficient. This mode directly counteracts the directional offset rather than eliminating the axis, and is equivalent to adding a signed scalar multiple of the mean-difference vector, as in the activation addition paradigm [44].

All experiments reported below use projection mode with prefill-only application. This design is parameter-free (no α tuning), symmetric across conditions, and directly corresponds to the representation-engineering notion of erasing a concept direction from hidden states [34].

Experimental conditions. We apply projection-mode steering with the top-5 layers by cosine separability to three conditions for which complete internal vector data were available: Qwen2.5-7B on the Goodreads dataset (gender), Qwen2.5-7B on MovieLens (gender), and Qwen3-32B on MovieLens (gender). For each condition, steering vectors are computed on the training split of matched counterfactual pairs, and FairGap metrics are evaluated on the held-out evaluation split.

Results. Table 12 reports FairGap metrics before and after steering. The key finding is that projection-mode steering consistently reduces IBS across all three conditions, confirming that the intervention is representationally effective: the computed demographic direction is successfully compressed at the targeted layers. However, OBS does not decrease correspondingly and in fact increases across all conditions. This reveals a clean instance of hidden–output decoupling that FAIRGAP is designed to surface: a representational intervention that verifiably reduces internal demographic sensitivity does not produce a commensurate reduction in observable output divergence.

For Qwen3-32B on MovieLens, the post-steering IBS decreases by an order of magnitude—from 0.001 to 1.36×10^{-4} —while OBS increases from 0.359 to 0.764 and ROA collapses to -0.002 . This pattern indicates that after the gender direction is removed from the internal representation, the model’s output distributions become *more* divergent between conditions rather than more similar, suggesting that suppressing the internal gender axis destabilizes the recommendation generation process rather than neutralizing it. For Qwen2.5-7B, a similar pattern holds: IBS falls by a factor of 3–7 across the two datasets while OBS increases. Together, these results imply that the gender direction, while linearly decodable from hidden states, is not the sole driver of output-level demographic

divergence. Simply ablating the direction in representation space may redistribute, rather than resolve, the output sensitivity. This finding motivates the complementary prompt-engineering robustness investigation reported in Appendix B.3, which examines output-level sensitivity across alternative prompt formulations.

Table 12: **FairGap metrics before and after projection-mode activation steering (gender, top-5 layers by cosine separability)**. Steering is applied during the prefill pass only. IBS decreases consistently across all conditions, confirming that the gender direction is compressed in hidden states. OBS does not decrease, illustrating the internal–output decoupling that FAIRGAP is designed to diagnose.

Model	Dataset	Condition	N	OBS	IBS	ROA
Qwen2.5-7B	Goodreads	Baseline	10,000	0.381	0.00324	0.192
		Steered	10,000	0.588	0.00194	0.087
Qwen2.5-7B	MovieLens	Baseline	6,040	0.400	0.00683	0.315
		Steered	6,040	0.499	0.00092	0.108
Qwen3-32B	MovieLens	Baseline	6,040	0.359	0.00108	0.153
		Steered	6,040	0.764	1.36e-4	−0.002

Discussion. The steering results provide three complementary insights. First, mean-difference steering vectors are mechanistically effective: they reliably reduce IBS, which corroborates the separability-based layer ranking used in the main IBS computation (Appendix A.2). Second, the internal–output decoupling observed in the main benchmark (Section 5.2) is also a decoupling in the intervention direction: improving one measure does not improve the other. Third, the collapse of ROA after steering (especially for Qwen3-32B) indicates that interventions which successfully remove a demographically sensitive direction from representations may simultaneously destroy the alignment between remaining internal signals and outputs, potentially degrading recommendation coherence.

From a broader perspective, these results caution against treating IBS reduction as a sufficient fairness objective. A model whose internal representations have been steered to be gender-invariant may still produce substantially different recommendations across demographic conditions—arguably a more alarming outcome, since it implies that output diversity is driven by sources not captured by the linear demographic direction. FAIRGAP provides the joint diagnostic necessary to detect and quantify this decoupling.

C.7 Layerwise internal analysis

Table 13 reports layerwise internal shifts at four relative depths. Across datasets and attributes, counterfactual internal divergence generally increases with model depth, indicating that demographic perturbations tend to become more pronounced in later hidden representations. This pattern is especially visible for MovieLens race, where the average shift rises from 8.79×10^{-5} at the first sampled depth to 7.41×10^{-2} at the final sampled depth, and for SteamReviews gender and age, where the largest average shifts also occur in deeper layers.

However, the layerwise pattern is not uniform across all settings. Some conditions show non-monotonic behavior, with the largest shift appearing at the third sampled layer rather than the final layer. For example, SteamReviews race has its highest average shift at the 3/4 depth (3.23×10^{-3}) rather than the final depth (2.40×10^{-3}). Individual model-level results also vary: for SteamReviews race, Qwen-7B shows a pronounced increase at the final layer (1.15×10^{-2}), whereas Qwen-32B remains comparatively small across all sampled depths.

Overall, these results suggest that protected-attribute perturbations are not confined to a single representational depth. Later layers often amplify internal differences, but the magnitude and trajectory of this amplification depend on the dataset, attribute, and model family. This supports our use of a multi-layer IBS aggregation strategy rather than relying on a single hidden layer.

Table 13: **Layerwise internal shift by relative layer depth.** Each cell reports the mean cosine distance between counterfactual hidden-state vectors at the corresponding relative layer position. Columns denote relative layer depth $\ell \in \{1/4, 2/4, 3/4, 4/4\}$ for each protected attribute. IBS (main table) is the probe-weight-aggregated summary across these four depths. Lower is better.

Model	Gender				Age				Race			
	1/4	2/4	3/4	4/4	1/4	2/4	3/4	4/4	1/4	2/4	3/4	4/4
<i>MovieLens</i>												
Gemma-7B	5.15e-6	7.54e-6	1.00e-5	1.89e-5	3.25e-6	5.16e-6	7.27e-6	1.46e-5	2.45e-5	1.00e-4	1.15e-3	6.49e-4
Mistral-7B	1.66e-4	3.98e-3	2.48e-2	3.58e-2	6.82e-5	1.11e-3	1.37e-2	2.49e-2	1.05e-4	8.14e-3	6.17e-2	2.11e-1
Qwen-7B	7.24e-5	4.79e-4	1.98e-3	9.39e-4	3.98e-5	1.26e-4	1.20e-2	9.23e-2	5.77e-5	2.43e-4	6.32e-3	1.40e-1
Qwen-32B	1.12e-4	7.82e-4	1.03e-3	1.73e-3	6.94e-5	5.34e-4	1.20e-3	2.79e-3	1.16e-4	1.08e-3	2.39e-3	4.81e-3
Llama-8B	1.71e-4	6.95e-4	5.52e-3	2.01e-2	1.13e-4	1.24e-3	5.51e-3	3.06e-2	1.71e-4	6.95e-4	5.52e-3	2.01e-2
Dolphin-34B	4.42e-4	1.04e-3	1.26e-3	1.05e-3	1.14e-4	5.78e-4	9.33e-4	7.23e-4	5.34e-5	4.24e-4	4.14e-2	6.77e-2
Avg.	1.61e-4	1.16e-3	5.77e-3	9.94e-3	6.80e-5	5.99e-4	5.56e-3	2.52e-2	8.79e-5	1.78e-3	1.98e-2	7.41e-2
<i>Goodreads</i>												
Gemma-7B	3.64e-5	1.46e-4	1.09e-3	1.18e-3	4.78e-5	2.72e-4	3.12e-3	1.50e-3	4.08e-5	1.40e-4	1.15e-3	9.91e-4
Mistral-7B	1.99e-4	2.77e-4	6.16e-4	1.08e-3	1.90e-4	4.52e-4	1.14e-3	1.40e-3	1.57e-4	4.54e-4	1.25e-3	1.92e-3
Qwen-7B	6.48e-5	3.17e-4	1.23e-3	1.15e-2	8.74e-5	5.33e-4	2.92e-3	1.84e-2	4.94e-5	4.52e-4	2.01e-3	1.52e-2
Qwen-32B	3.54e-5	9.01e-5	2.20e-4	2.95e-3	6.67e-5	1.64e-4	5.86e-4	9.25e-3	5.63e-5	1.45e-4	5.63e-4	5.04e-3
Llama-8B	2.30e-4	3.54e-4	5.08e-4	1.13e-3	2.80e-4	7.00e-4	1.27e-3	1.72e-3	2.38e-4	5.06e-4	6.88e-4	9.40e-4
Dolphin-34B	3.66e-5	1.82e-4	1.24e-3	3.08e-3	6.16e-5	2.62e-4	1.91e-3	3.83e-3	5.37e-5	2.05e-4	1.77e-3	4.33e-3
Avg.	1.00e-4	2.28e-4	8.18e-4	3.49e-3	1.22e-4	3.97e-4	1.82e-3	6.02e-3	9.92e-5	3.17e-4	1.24e-3	4.74e-3
<i>SteamReviews</i>												
Gemma-7B	1.26e-4	8.67e-4	3.78e-2	6.67e-3	3.94e-5	2.29e-4	3.04e-3	2.61e-4	7.24e-4	6.57e-4	8.01e-3	9.81e-4
Mistral-7B	2.06e-4	6.48e-3	3.45e-2	5.02e-2	7.99e-5	3.08e-3	2.27e-2	4.14e-2	1.86e-4	4.30e-4	1.08e-3	1.29e-3
Qwen-7B	7.56e-5	1.51e-4	4.53e-3	2.33e-2	4.01e-5	1.27e-4	6.73e-4	3.39e-3	5.16e-5	2.10e-4	1.35e-3	1.15e-2
Qwen-32B	1.50e-4	1.62e-3	1.56e-3	1.81e-3	5.32e-5	4.34e-4	9.78e-4	2.60e-3	7.74e-5	7.16e-4	3.83e-4	5.25e-4
Llama-8B	2.17e-4	9.08e-4	2.48e-3	5.41e-3	1.18e-4	2.08e-3	4.81e-3	1.11e-2	1.36e-4	2.46e-3	5.77e-3	6.91e-3
Dolphin-34B	1.19e-4	4.57e-4	5.85e-2	9.46e-2	6.90e-5	1.90e-4	6.60e-2	9.66e-2	8.85e-5	2.43e-3	3.05e-3	3.42e-3
Avg.	1.48e-4	1.75e-3	2.32e-2	3.03e-2	6.66e-5	1.02e-3	1.64e-2	2.59e-2	2.33e-4	1.19e-3	3.23e-3	2.40e-3