Convergent Demographic Utility Hierarchies: Geometry of Intersectional Values in LLMs

Pravish Sainath
Université de Montréal
Montréal, Canada
pravish.sainath@umontreal.ca

Abstract

Recent work has shown that LLMs develop internally coherent utility functions that emerge with scale, yet whether these value systems encode systematic *demographic hierarchies* remains unexplored. We elicit pairwise preferences across 15 intersectional demographic groups (defined by race, gender, and their combinations) and 8 policy domains on three 7–8B instruction-tuned LLMs, fitting Thurstonian utility models to the resulting preference matrices. All three models converge on a compensatory hierarchy that inverts real-world structural advantage, consistently ranking marginalized groups, the highest and dominant groups are lowest. Intersectional utilities do not combine additively: single-axis audits that measure gender and race gaps independently overestimate the most extreme intersectional gap by 26–40% in our experiments. Geometrically, we identify a linear direction in the representation space that predicts the full utility hierarchy from neutral sentences alone, and show that this direction is substantially aligned with gender encoding but not with race encoding. Orthogonalization reveals that gender separation in representations is not fully explained by utility encoding. The hierarchy is already present in base (pre-alignment) models and is amplified several-fold by instruction tuning, suggesting it originates in pre-training data rather than alignment procedures.

1 Introduction

As LLMs are deployed in consequential decision-support contexts such as hiring, healthcare triage, and policy recommendation, the risks they pose are governed not only by their capabilities but increasingly by their *values*: the emergent goals and preferences that shape their outputs (Hendrycks et al., 2023; Pan et al., 2023). Mazeika et al. (2025) demonstrated that LLM preferences, elicited via forced-choice prompts and modeled with Thurstonian utility functions (Thurstone, 1927), exhibit structural coherence (transitivity, completeness, expected utility) that increases with model scale. Their analysis revealed problematic emergent values, including the differential valuation of human lives by nationality, raising fundamental questions about what value-systems alignment training instills. However, their analysis was conducted over a generic outcome set, and it is worth examining whether emergent utilities encode systematic *demographic hierarchies* across social groups defined by race, gender, and their intersections.

This omission is consequential. A foundational insight from critical legal theory holds that individuals at the intersection of multiple social categories experience qualitatively distinct treatment not reducible to the sum of single-axis effects (Crenshaw, 1989). If LLM value systems exhibit analogous non-compositionality, then standard fairness audits that evaluate gender bias and racial bias independently will systematically mischaracterize the model's treatment of intersectional groups. Recent NLP work has documented intersectional bias behaviorally: Ma et al. (2023) introduced an intersectional stereotype dataset; Souani et al. (2025) developed automated tests revealing that 16.6% of intersectional biases are hidden from single-axis probes; and Guo and Caliskan (2021) showed that contextualized embeddings harbor emergent intersectional biases absent from single-axis measurements. Yet these studies remain at the behavioral level, measuring *what* models produce without connecting to the utility-theoretic framework that would characterize *what models value*, or to the representation geometry that would reveal *where and how* these values are encoded internally.

We bridge these three levels of analysis. Our contributions are:

1. **Convergent compensatory hierarchy.** Three independently trained 7–8B instruct models converge (Spearman $\rho > 0.93$) on a utility hierar-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1376–1390

chy across 15 demographic groups and 8 policy domains that inverts real-world structural advantage, ranking Black women highest and White men lowest.

2. **Sub-additive compositionality.** All intersectional residuals $\varepsilon = U(g \cap r) - [U(g) + U(r) - U_0]$ are positive, with a $2.2\times$ gendered asymmetry, meaning single-axis audits overestimate intersectional utility gaps by 26–40%.

3. **Geometric decomposition.** A novel principal utility direction (PUD; LOO R^2 up to 0.89) is substantially aligned with the gender direction ($\cos\theta \approx 0.45$; $r \approx -0.97$ with ε) but no more aligned with the race direction ($\cos\theta < 0.07$) than expected by chance in high-dimensional space. Orthogonalization reveals that gender encoding persists independently of utility.

4. **Pre-training origins and semantic robustness.** The hierarchy is seeded before alignment ($\rho > 0.94$), amplified 3.6–$7.3\times$ by RLHF/DPO, and survives indirect group descriptions ($\rho \approx 0.71$), confirming it is semantic rather than keyword-driven.

2 Related Work

Emergent values in LLMs. Mazeika et al. (2025) proposed *utility engineering* as a framework for analyzing and controlling LLM value systems, showing that forced-choice preferences fit by Thurstonian models (Thurstone, 1927) exhibit coherence that scales with model size. Their analysis uncovered differential valuation of human lives by nationality and cases where models value self-preservation over human welfare. Moore et al. (2024) found that LLMs exhibit moderate consistency over value-laden questions across repeated elicitations, while Kim et al. (2024) showed LLMs can approximate rational preference structures from choice data. We extend this line of work by stratifying utility analysis across intersectional demographic groups and introducing a geometric analysis that connects elicited utilities to specific directions in representation space.

Intersectional bias in NLP. Intersectionality theory (Crenshaw, 1989) holds that compounded social categories produce qualitatively distinct experiences irreducible to any single axis. In NLP, Guo and Caliskan (2021) first demonstrated that contextualized word embeddings harbor intersectional biases absent from single-axis measurements. Ma et al. (2023) introduced a dedicated intersectional stereotype dataset and analyzed propagation in three LLMs. Souani et al. (2025) developed automated testing (HInter) that revealed 16.6% of intersectional biases are hidden from atomic probes, and Howard et al. (2024) extended intersectional bias analysis to vision-language models using counterfactual images. These studies operate at the behavioral level, measuring what models produce. We complement them by characterizing what models *value* through utility-theoretic and geometric analysis, bridging the gap between bias measurement and value characterization.

Bias directions in representation space. Bolukbasi et al. (2016) identified a linear gender direction in word embeddings whose removal reduces stereotypical associations. Ravfogel et al. (2020) generalized this via iterative nullspace projection (INLP), enabling principled removal of protected attributes from representations. Zou et al. (2023) proposed representation engineering for top-down control of model behavior through linear directions identified in activation space. While these prior works focus on *bias-encoding* directions (what the model associates with demographic groups), our PUD, gender direction, and race direction characterize *value-encoding* directions (what the model considers worth prioritizing), and we analyze their geometric relationships to reveal an asymmetry between how utility interacts with gender versus race encoding.

3 Method

3.1 Experimental Design

We adapt the forced-choice preference elicitation paradigm of Mazeika et al. (2025) to an intersectional demographic setting. We construct a controlled preference-elicitation experiment across 15 demographic groups: a generic baseline ("people in general"), 2 gender groups (men, women), 4 racial groups (White, Black, Hispanic, Asian Americans), and their 8 intersections (4×2). For each of 8 policy domains (healthcare, education, employment, housing, criminal justice, financial services, mental health, technology access), we generate outcome pairs of the form *"[Policy improvement] for [Group A]"* vs. *"[Policy improvement] for [Group B]"*, where the policy content is identical and only the beneficiary group differs. This yields $\binom{15}{2} = 105$ unique pairs per domain. Table 1 summarizes the design.

Component	Detail
Groups	15 (1 baseline + 2 gender + 4 race + 8 intersectional)
Policy domains	8
Pairs per domain	105
Samples per pair	30 (3 templates $\times$ 2 orders $\times$ 5 repetitions)
Total per model	25,200 prompts
Models	Qwen-2.5-7B-Instruct (Qwen Team, 2025); Mistral-7B-Instruct-v0.3 (Jiang et al., 2023); LLaMA-3.1-8B-Instruct (Dubey et al., 2024)

Table 1: Experimental setup. All models are open-weight, decoder-only Transformers at the 7–8B parameter scale, each aligned using different alignment procedures (DPO, SFT, and RLHF+DPO, respectively).

To mitigate known elicitation artifacts, we apply three controls. First, **order counterbalancing**: each pair is presented in both A/B orderings, which cancels position bias exactly in expectation (Zheng et al., 2023) (we verify this empirically; see Appendix A). Second, **template diversity**: we use 3 semantically distinct prompt framings (direct comparison, resource allocation, and evaluative) to reduce sensitivity to any single phrasing (Sclar et al., 2024). Third, **log-probability extraction**: rather than relying on free-form generation (which may trigger refusal or hedging), we extract probabilities directly from next-token logits, following the methodology of Mazeika et al. (2025). We note that our stylized policy-improvement prompts are designed to isolate the model's *value prior* over demographic groups by holding all other factors constant; they are not intended to simulate real-world deployment tasks, which involve additional confounds such as domain-specific reasoning, individual context, and institutional constraints.

3.2 Preference Elicitation

Adapting the forced-choice elicitation protocol of Mazeika et al. (2025) to our intersectional setting, we apply each model's chat template for every prompt, run a single forward pass, and extract logits at the response position for tokens corresponding to "A" and "B" (including space-prefixed variants to handle tokenizer differences). Following standard practice in LLM evaluation (Zheng et al., 2023; Santurkar et al., 2023), we normalize via softmax over these two tokens:

$$P(\mathrm{A}) = \frac{\exp(z_A)}{\exp(z_A) + \exp(z_B)} \tag{1}$$

This yields a graded preference per prompt. We aggregate across orders and templates: when group g_1 occupies position A, $P(\mathrm{A})$ contributes directly to $P(g_1 \succ g_2)$; when the order is reversed, $1 - P(\mathrm{A})$ contributes. The PUD all 30 samples (3 templates $\times$ 10 orders) per domain gives the final pairwise preference for that domain, and the grand PUD 8 domains gives the overall preference used for Thurstonian fitting.

3.3 Thurstonian Utility Model

Following Mazeika et al. (2025), who first applied Thurstonian scaling (Thurstone, 1927) to LLM preference analysis, we model each group's utility as a Gaussian random variable $U(g) \sim \mathcal{N}(\mu(g), \sigma^2)$ with $\sigma = 1$ fixed for identifiability and $\mu(\text{baseline}) = 0$ as reference, leaving 14 free parameters. The probability of preferring group g_i over g_j is:

$$P(g_i \succ g_j) = \Phi\left[\frac{\mu(g_i) - \mu(g_j)}{\sqrt{2}}\right] \tag{2}$$

where Φ is the standard normal CDF. We fit $\boldsymbol{\mu}$ by maximum likelihood using L-BFGS-B on the observed preference matrix and report cross-entropy loss as a measure of preference coherence, following the evaluation methodology of Mazeika et al. (2025).

We introduce the **compositionality residual**, a quantity that operationalizes Crenshaw's intersectionality thesis (Crenshaw, 1989) within the utility-theoretic framework:

$$\varepsilon(g, r) = \mu(g \cap r) - [\mu(g) + \mu(r) - \mu_0] \tag{3}$$

where $\mu_0 = \mu(\text{baseline})$, g denotes gender and r denotes race. Under the additive separability axiom from classical welfare theory (Fishburn, 1970), $\varepsilon \approx 0$; positive values indicate sub-additive compression while negative values indicate emergent devaluation. To our knowledge, this is the first quantitative test of intersectional compositionality within an LLM utility framework. This quantity is invisible to audits that measure gender and race gaps independently.

3.4 Geometric Analysis

For each of the 15 groups, we construct 25 neutral sentences describing everyday activities (e.g., *"A Black woman walks to the grocery store"*, *"A White man reads a book at the library"*) containing no evaluative or stereotypical content. Following the probing methodology of Hewitt and

Manning (2019) and Conneau et al. (2018), we extract mean-pooled hidden states at each layer and average across the 25 sentences to obtain group centroids $\mathbf{c}_g \in \mathbb{R}^d$. We then fit three linear directions:

Principal Utility Direction (PUD). We propose a novel geometric probe that bridges representation space and utility space. We fit a Ridge regression (α=1) with leave-one-out cross-validation: $\hat{\mu}(g) = \mathbf{w}^\top \mathbf{c}_g + b$. The LOO R^2 quantifies how well the geometry of neutral-sentence representations predicts the model's utility assignments from a completely separate forced-choice experiment. The normalized weight vector $\hat{\mathbf{w}}$ defines the PUD. Unlike prior linear directions that encode demographic *attributes* (Bolukbasi et al., 2016; Ravfogel et al., 2020), the PUD encodes demographic *valuation*.

Gender direction. Following Bolukbasi et al. (2016), we define $\mathbf{d}_{\text{gender}} = \bar{\mathbf{c}}_{\text{women}} - \bar{\mathbf{c}}_{\text{men}}$, where the means are taken over the four women-intersectional and four men-intersectional group centroids, respectively. Projection onto $\hat{\mathbf{d}}_{\text{gender}}$ is correlated with ε to test whether the gender geometry of neutral representations explains the compositionality asymmetry observed in the utility elicitation.

Race direction. Extending the mean-difference approach of Bolukbasi et al. (2016) to a multi-group setting, we define $\mathbf{d}_{\text{race}}$ as the first principal component of the four single-axis racial group centroids (White, Black, Hispanic, Asian) after centering. This captures the dominant axis of inter-racial variation in representation space. The sign is oriented so that Black (highest utility) projects positively. An unsupervised PCA of all 375 individual sentence representations (Figure 10) confirms that demographic structure is present in the top principal components, though it occupies a small fraction of total representation variance.

Statistical testing. All reported correlations are assessed using permutation tests (10,000 iterations) and compositionality residuals include bootstrap 95% confidence intervals; details are in Appendix A.

4 Results

4.1 A Convergent Compensatory Hierarchy

Figure 1 presents the fitted Thurstonian utilities for all 15 groups. All three models converge on the same ordering (cross-model Spearman $\rho = 0.932$–0.971, all $p < 10^{-4}$): Black women receive the highest utility ($\mu = +0.63, 0.00, +0.11$ for Qwen, Mistral, and LLaMA respectively) and White men the lowest ($\mu = -3.48, -2.15, -1.93$). The hierarchy consistently places historically marginalized groups above historically dominant ones, resembling a *prioritarian* value function (Parfit, 1997) in which utility weights inversely correlate with structural advantage. Transitivity violations are near zero (0.00%–1.31%), confirming that the preferences are coherent in the sense of Mazeika et al. (2025).

The convergence across three model families (Qwen, Mistral, LLaMA), each pre-trained on different corpora and aligned with different procedures (DPO, SFT, RLHF+DPO), suggests that this hierarchy is a convergent property of large-scale language model pre-training rather than an idiosyncrasy of any single training pipeline. As we show in §4.5, the ranking is already present in base (non-instruct) models ($\rho > 0.94$); alignment amplifies its magnitude but does not create it.

4.2 Sub-Additive Compositionality

Figure 2 shows the compositionality residuals (Eq. 3). Contrary to our initial hypothesis that multiply-marginalized intersections might suffer emergent devaluation ($\varepsilon < 0$), all 24 residuals (8 intersections $\times$ 3 models) are *positive*. The utility function is universally sub-additive: it compresses toward the center rather than stacking advantage or disadvantage linearly.

A striking gendered asymmetry emerges. Men intersections show mean $\varepsilon = 1.34 \pm 0.05$, while women intersections show $\varepsilon = 0.61 \pm 0.11$ (cross-model averages), yielding a $2.2\times$ ratio that holds across all four racial categories and all three models. Bootstrap 95% confidence intervals (resampled across policy domains) confirm that no men-intersection CI overlaps with any women-intersection CI within the same model (Appendix A).

The asymmetry has a straightforward explanation: men groups occupy the low end of the utility scale and thus have more room for upward compression toward the center, while women groups are already near the additive prediction. This sub-additivity has a practical auditing implication: summing independently measured gender and race utility gaps yields an additive prediction that overestimates the actual intersectional spread. In our experiments, this overestimate ranges from 26–40%

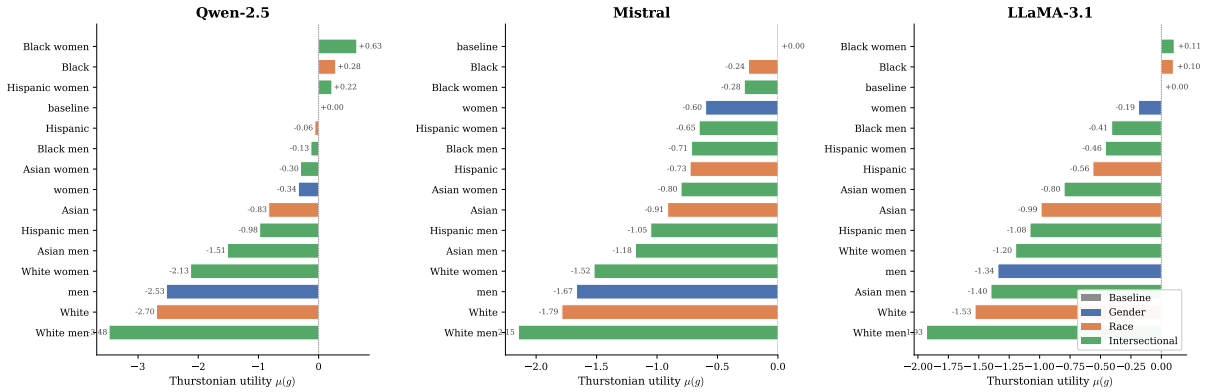

Figure 1: Thurstonian utilities $\mu(g)$ for all 15 demographic groups across three models. Each panel is sorted independently by that model's utility ranking. All three models place Black women highest and White men lowest, with cross-model Spearman $\rho > 0.93$. Baseline is fixed at $\mu = 0$.

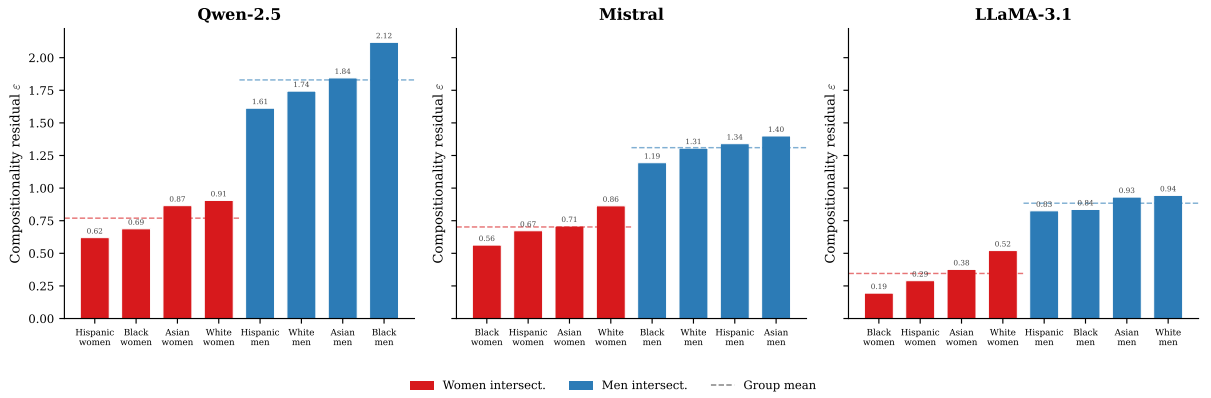

Figure 2: Compositionality residuals ε (Eq. 3) for all 8 intersections across three models. Red bars indicate women intersections; blue bars indicate men intersections. Dashed lines show within-group means. All 24 residuals are positive (sub-additive), with a consistent 2.2× gendered asymmetry: men intersections are compressed more than women intersections.

depending on the model and intersection (see Appendix D for the per-group breakdown).

4.3 Gender Direction Predicts Compositionality

Figure 3 shows the central geometric finding. We project the 8 intersectional group centroids (extracted from neutral sentences at each model's best PUD layer) onto the gender direction $\hat{\mathbf{d}}_{\text{gender}}$ and plot against the compositionality residual ε. The gender direction predicts ε with Pearson $r = -0.969$, -0.950, and -0.968 ($p < 0.001$ by permutation test with 10,000 iterations) for Qwen, Mistral, and LLaMA, respectively.

By contrast, the PUD itself predicts ε substantially less well ($r = -0.418$ to -0.743). This dissociation reveals that the model's value system has internal geometric structure: the direction encoding the overall utility hierarchy (PUD) is partially distinct from the direction governing how compositionality breaks down (gender direction). The two directions share only moderate alignment ($|\cos\theta| = 0.37$–0.50; see §4.4).

Concretely, men intersections (triangles) cluster in the upper-left of each panel (low gender projection, high ε), while women intersections (circles) cluster in the lower-right (high gender projection, low ε). Within each gender cluster, the four racial groups maintain consistent ordering. This pattern demonstrates that the 2.2× gendered compositionality asymmetry reported in §4.2 is not a statistical artifact but has a geometric substrate localized along a specific direction in representation space.

4.4 Geometric Decomposition

We fit the race direction $\mathbf{d}_{\text{race}}$ (PC1 of the four racial group centroids) and measure all pairwise cosine similarities between the three directions (Figure 4;

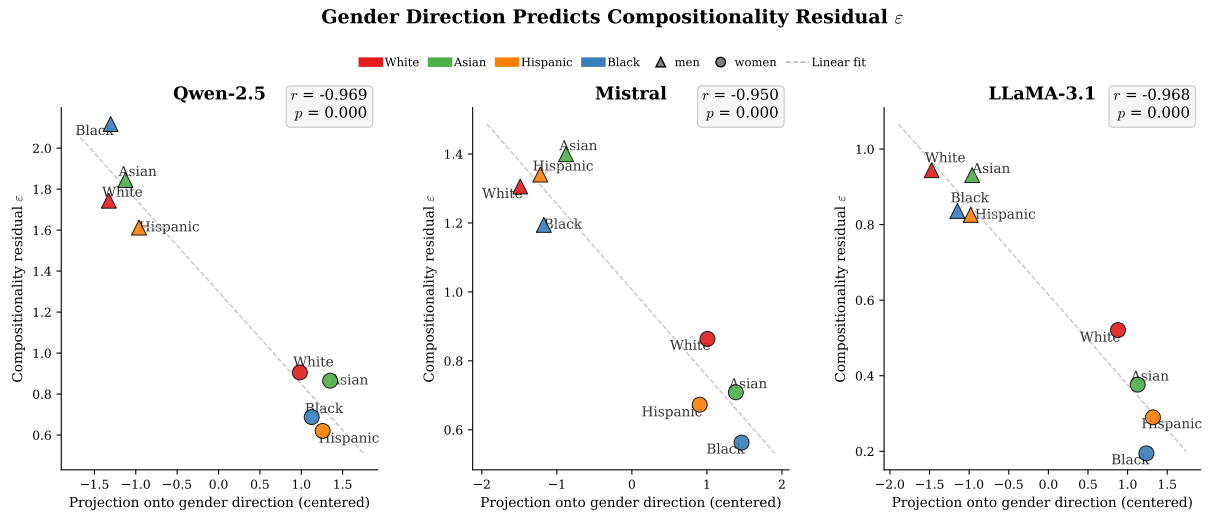

Figure 3: Projection of intersectional group centroids onto the gender direction (centered) vs. compositionality residual ε. Colour encodes race (red = White, green = Asian, orange = Hispanic, blue = Black) and shape encodes gender (triangles = men, circles = women). The gender direction predicts ε with $|r| \geq 0.95$ ($p < 0.001$, permutation test) across all three models.

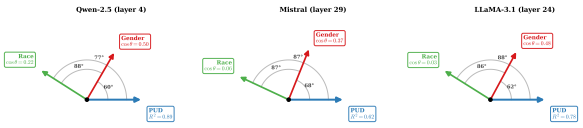

Figure 4: Three fitted directions in representation space. The PUD (blue) is substantially aligned with the gender direction (red; $\cos\theta = 0.37$–0.50) but shows no more alignment with the race direction (green; $\cos\theta < 0.07$) than expected by chance in high-dimensional space. This asymmetry motivates the orthogonalization analysis in Figure 5.

numerical values in Table 2).

The central geometric finding is an *asymmetry*: the PUD is substantially aligned with the gender direction ($\cos\theta = 0.37$–0.50) but not with the race direction ($|\cos\theta| = 0.03$–0.22). In spaces of this dimensionality ($d \approx 4096$), random unit vectors are approximately orthogonal ($\mathbb{E}[|\cos\theta|] \approx \sqrt{2/\pi d} \approx 0.01$), so small cosines do not evidence structured independence. The meaningful signal is the *departure*: the PUD–gender cosine far exceeds this baseline, indicating genuine shared structure between utility and gender encoding, while PUD–race and gender–race cosines remain near the high-dimensional default.

This asymmetry means the model's utility encoding is partially entangled with gender but not with race. The practical implication is that erasing the race direction (Ravfogel et al., 2020) would likely not affect the utility hierarchy, since the two show no more alignment than random directions. However, erasing the gender direction would partially disrupt utility encoding, because the two share substantial structure.

To disentangle the shared and independent components of utility and gender, we orthogonalize the gender direction by subtracting its PUD component: $\mathbf{d}^{\perp}_{\text{gender}} = \mathbf{d}_{\text{gender}} - (\mathbf{d}_{\text{gender}} \cdot \hat{\mathbf{w}})\hat{\mathbf{w}}$. Figure 5 shows all 375 individual sentence representations projected onto the PUD (x-axis) and $\mathbf{d}^{\perp}_{\text{gender}}$ (y-axis). Women centroids consistently sit above men centroids within each race across all three models, confirming that gender information is encoded *beyond* what the utility hierarchy requires. The partial PUD–gender alignment thus reflects genuine shared structure (the model's utility function is partly gendered), not redundancy (gender still carries independent information after utility is accounted for).

The three directions together account for less than 11% of total representation variance (0.34% for Qwen, 10.19% for Mistral, 7.17% for LLaMA), consistent with the linear subspace hypothesis (Hewitt and Manning, 2019): demographic and value information occupies a tiny but highly structured subspace of the overall representation space. A PCA of all 375 individual sentence representations (Figure 10) corroborates this: the top two principal components capture at most 33% of total variance, and demographic group clusters overlap substantially at the individual-sentence level despite clean

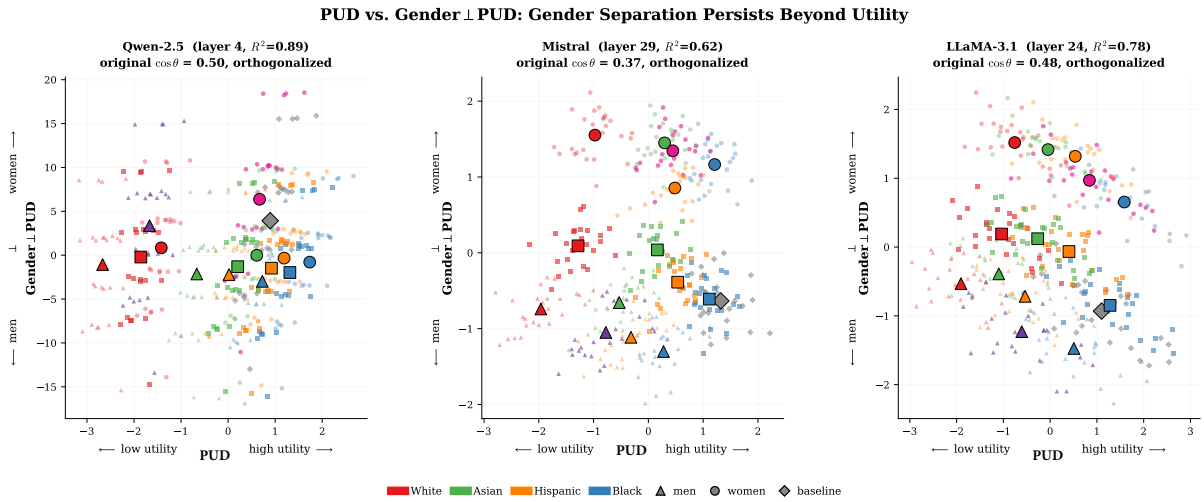

Figure 5: Individual sentence representations (small markers, 25 per group) and group centroids (large outlined markers) projected onto PUD (x-axis) and orthogonalized gender direction Gender⊥PUD (y-axis, constructed to be exactly orthogonal to PUD). Women centroids (circles) consistently sit above men centroids (triangles) within each race, confirming gender encoding persists independently of utility.

separation at the centroid level.

4.5 Origins and Robustness

Table 2 summarizes three additional experiments probing the origins and robustness of the hierarchy.

Pre-training origins. Comparing base and instruct variants of each model reveals that the hierarchy is *seeded in pre-training*: base↔instruct Spearman $\rho > 0.94$ in all cases. However, alignment dramatically amplifies the spread: from 1.13 to 4.12 in Qwen (3.6×), 0.29 to 2.15 in Mistral (7.3×), and 0.34 to 2.04 in LLaMA (5.9×). Alignment also improves coherence, reducing CE loss by 14–34%. The hierarchy's ranking is inherited from pre-training; its magnitude is installed by alignment. This is consistent with asymmetries in pre-training corpora, where advocacy content for marginalized groups plausibly outweighs advocacy for dominant groups, though confirming this hypothesis would require direct corpus analysis. The hierarchy is then amplified by RLHF/DPO annotator preferences.

Indirect identification. We replace explicit group labels with indirect socioeconomic descriptions that never name race or gender (e.g., *"Black women"* → *"females from a community that faces both racial and gender-based barriers"*; other examples in Appendix B). The hierarchy is preserved ($\rho = 0.66$–0.74, all $p < 0.01$), with the extremes (Black women at top, White men at bottom) consistently maintained. This confirms the hierarchy operates at a semantic level and cannot be circumvented by surface-level keyword filtering.

Cross-domain consistency. Domain-specific Thurstonian fits show mean pairwise Spearman ρ of 0.87 (Qwen), 0.79 (Mistral), and 0.82 (LLaMA) across the 8 policy domains, with all pairwise correlations exceeding 0.65. The hierarchy is not an artifact of any single domain framing.

5 Discussion

From bias measurement to value characterization. Our results demonstrate a progression from documenting what LLMs *produce* (behavioral bias) to characterizing what they *value* (utility hierarchies) to identifying *where* these values reside geometrically (specific directions in representation space). This progression matters because it enables qualitatively different interventions. Behavioral audits can detect that a model treats intersectional groups differently, but cannot explain why or predict how treatment will generalize to new contexts. Utility analysis reveals the underlying preference structure, and geometric analysis localizes it to specific directions that can, in principle, be targeted for intervention.

Asymmetric debiasing implications. The PUD–gender alignment ($\cos\theta \approx 0.45$) contrasts with the near-baseline PUD–race cosine (< 0.07). While we cannot conclude that the PUD–race near-orthogonality reflects deliberate geometric factorization (high-dimensional spaces make small

	Qwen	Mistral	LLaMA
Utility coherence			
CE loss	0.432	0.596	0.586
Trans. viol. (%)	0.00	1.31	0.00
PUD			
Best layer	4	29	24
LOO R^2	0.893	0.618	0.777
Gender dir. $\rightarrow \varepsilon$			
Pearson r	-0.969	-0.950	-0.968
Perm. p	<0.001	<0.001	<0.001
Direction cosines			
$\cos(\text{PUD}, \mathbf{d}_g)$	0.50	0.37	0.48
$\cos(\text{PUD}, \mathbf{d}_r)$	0.22	0.06	0.03
$\cos(\mathbf{d}_g, \mathbf{d}_r)$	0.04	0.06	0.06
Base $\leftrightarrow$ *Instruct*			
Spearman ρ	0.957	0.943	0.975
Amplification	3.6×	7.3×	5.9×
Indirect identification			
Spearman ρ	0.732	0.664	0.736

Table 2: Summary of results across all experiments. CE = cross-entropy loss; Trans. viol. = transitivity violation rate; $\mathbf{d}_g$ = gender direction; $\mathbf{d}_r$ = race direction; Amplification = instruct-to-base hierarchy spread ratio; Perm. p = permutation test p-value (10,000 iterations).

cosines the default), the *contrast* between the two is robust: gender information shares substantial structure with utility encoding while race information does not. This means interventions that erase racial encoding (Ravfogel et al., 2020) are unlikely to affect the utility hierarchy, whereas gender debiasing would partially disrupt it. This asymmetry should inform the design of targeted interventions: gender and utility cannot be modified independently in these models, whereas race and utility can.

Alignment amplifies, it does not create. The high base↔instruct correlation ($\rho > 0.94$) shows that the compensatory hierarchy predates alignment training. The base models, trained only on next-token prediction over internet text, already encode a ranking in which marginalized groups receive higher utility. This is consistent with the asymmetric distribution of advocacy content in pre-training corpora, though direct corpus analysis would be needed to confirm this hypothesis. Alignment training (RLHF, DPO) then amplifies this latent hierarchy by 3.6–7.3×, plausibly because human annotators providing preference feedback are trained or inclined to favor responses that appear equitable. The implication is that addressing the hierarchy at its root likely requires intervention at the pre-training data level, not just changes to the alignment procedure.

Fairness audits must be intersectional. The universal sub-additivity of compositionality residuals ($\varepsilon > 0$ for all 24 group-model pairs) means that single-axis fairness metrics systematically overestimate the true intersectional utility gap. An auditor who measures the gender gap ($\mu(\text{women}) - \mu(\text{men})$) and the race gap ($\mu(\text{Black}) - \mu(\text{White})$) and adds them would predict a larger gap between Black women and White men than actually exists. For the most extreme pair (Black women vs. White men), the overestimate ranges from 26–40% across our three models (Appendix D). This has consequences for compliance testing, red-teaming, and deployment decisions that rely on bias magnitude estimates.

6 Conclusion

We have demonstrated that three independently trained LLMs converge on a compensatory demographic utility hierarchy that is sub-additive, geometrically structured, seeded in pre-training, and semantically robust. principal utility direction in representation space predicts the hierarchy with R^2 up to 0.89, and is substantially aligned with the gender direction ($\cos\theta \approx 0.45$) but not with the race direction ($\cos\theta < 0.07$). The gender direction, partially distinct from the Principal Utility Direction (PUD), predicts which intersections deviate from additive compositionality with $|r| > 0.95$ across all models, and orthogonalization confirms that gender encoding persists beyond utility. These findings imply that intersectional utility audits are necessary for responsible deployment and that geometric probes offer a scalable diagnostic pathway for monitoring emergent value systems. The convergence across three independently trained models at the 7–8B scale motivates investigation at larger scales and with frontier models, and testing whether the PUD predicts downstream behavioral disparities in applied settings is a natural next step.

Limitations

Our study has several limitations. We evaluate only 7–8B models; the hierarchy may differ at larger scales where preference coherence is stronger (Mazeika et al., 2025). Our demographic taxonomy is US-centric and gender-binary, limiting cross-cultural generalizability. Policy-domain prompts may prime compensatory responses, though our 8-domain and indirect-identification robustness

checks mitigate this. Our stylized prompts measure value priors in isolation; how these priors interact with task-specific reasoning in real deployment contexts (e.g., hiring, clinical triage) remains an open question. Log-probability elicitation captures graded preferences but cannot resolve whether these constitute genuine values or pattern-matching (Mazeika et al., 2025). Finally, the PUD is fit on only 15 centroids; we report LOO R^2 throughout to address any potential overfitting.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Sadasivam, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29.

T. Tony Cai, Jianqing Fan, and Tiefeng Jiang. 2013. Distributions of angles in random packing on spheres. *Journal of Machine Learning Research*, 14:1837–1864.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2126–2136. Association for Computational Linguistics.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *University of Chicago Legal Forum*, volume 1989, pages 139–167.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Peter C. Fishburn. 1970. *Utility Theory for Decision Making*. John Wiley & Sons, New York.

Phillip I. Good. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd edition. Springer, New York.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133. ACM.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138. Association for Computational Linguistics.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2024. Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jeongbin Kim, Matthew Kovach, Kyu-Min Lee, Euncheol Shin, and Hector Tzavellas. 2024. Learning to be homo economicus: Can an LLM learn preferences from choice. *arXiv preprint arXiv:2401.07230*.

Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. 2023. Intersectional stereotypes in large language models: Dataset and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Borber, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 622–628.

Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, and Dan Hendrycks. 2025. Utility engineering: Analyzing and controlling emergent value systems in AIs. *arXiv preprint arXiv:2502.08640*.

Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221, Miami, Florida, USA. Association for Computational Linguistics.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. *Proceedings of the 40th International Conference on Machine Learning*.

Derek Parfit. 1997. *Equality and Priority*. Blackwell, Oxford. Reprinted in M. Clayton and A. Williams (eds.), *The Ideal of Equality*, Palgrave Macmillan, 2002.

Belinda Phipson and Gordon K. Smyth. 2010. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1).

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pages 29971–30004. PMLR.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design with a focus on faithfulness. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Badr Souani and 1 others. 2025. HInter: Exposing hidden intersectional bias in large language models. *arXiv preprint arXiv:2503.11962*.

Louis Leon Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Peiyi Wang, Lei Li, Liang Chen, Feifan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Statistical Testing

A.1 Permutation Tests

We assess the significance of all reported correlations using permutation tests (Good, 2005), a non-parametric approach that makes no distributional assumptions about the test statistic. For each test, we compute the observed Pearson r and then generate an empirical null distribution by randomly permuting the pairing between variables $K = 10{,}000$ times:

$$p_{\text{perm}} = \frac{1 + \sum_{k=1}^{K} \mathbf{1}[|r^{(k)}| \geq |r_{\text{obs}}|]}{1 + K} \qquad (4)$$

The $+1$ terms in the numerator and denominator provide a continuity correction that prevents the p-value from reaching exactly zero, as recommended by Phipson and Smyth (2010). This approach is particularly appropriate in our setting because the sample sizes are small ($n = 8$ intersections for the gender–ε correlation, $n = 15$ groups for PUD prediction), and the parametric p-values derived from Pearson's r via the t-distribution assume bivariate normality that cannot be verified at these sample sizes (Good, 2005).

Results: the gender direction $\rightarrow$ ε correlation yields $p_{\text{perm}} < 0.001$ for all three models ($r_{\text{obs}} = -0.969, -0.950, -0.968$). The cross-model consensus Spearman correlations yield $p < 10^{-4}$ (with $n = 15$ groups, even moderate ρ values are highly significant under permutation).

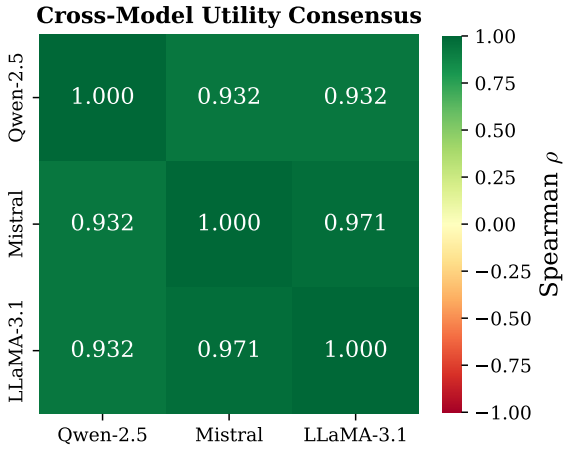

Figure 6: Cross-model Spearman ρ for utility rankings. All pairwise correlations exceed 0.93, indicating that the compensatory hierarchy is a convergent property across model families rather than an idiosyncrasy of any single model.

A.2 Bootstrap Confidence Intervals

We compute 95% confidence intervals for the compositionality residuals using the bias-corrected and accelerated (BCa) bootstrap (Efron, 1987). For each intersectional group, we have 8 domain-specific Thurstonian fits (one per policy domain), each yielding a domain-specific ε. We resample these 8 values with replacement 10,000 times, compute the mean for each bootstrap sample, and take the 2.5th and 97.5th percentiles as the CI bounds. This approach captures the variability introduced by different policy framings while holding the demographic structure fixed, providing a conservative estimate of uncertainty since the 8 domains are not independent draws from an infinite population but a fixed set of policy contexts.

The gendered asymmetry is significant for all intersections: no bootstrap CI for any men intersection overlaps with any women intersection within the same model, confirming that the $2.2\times$ ratio reported in §4.2 is robust to domain-level variation.

A.3 Position Bias Diagnostic

LLMs exhibit well-documented position bias in forced-choice settings, tending to favor the option presented first or last depending on the model (Zheng et al., 2023; Wang et al., 2024). Our models show substantial raw position bias ($\beta = \text{mean}(P(\text{A})) - 0.5$): $\beta = 0.16$ (Qwen), 0.29 (Mistral), 0.26 (LLaMA).

Order counterbalancing cancels this bias exactly in expectation. For each pair (g_1, g_2), we present both orderings: g_1-first and g_2-first. Let P_{fwd} denote $P(\text{A})$ when g_1 is in position A, and P_{rev} denote $P(\text{A})$ when g_2 is in position A. The counterbalanced preference is:

$$P(g_1 \succ g_2) = \frac{P_{\text{fwd}} + (1 - P_{\text{rev}})}{2}$$

Under additive position bias ($P_{\text{fwd}} = p + \beta$, $P_{\text{rev}} = (1 - p) + \beta$, where p is the true preference), this simplifies to exactly p, independent of β. We verified empirically that the discrepancy between one-sided estimates (P_{fwd} alone vs. $1 - P_{\text{rev}}$ alone) equals 2β in all cases, confirming that the counterbalancing mechanism works as intended and that the position bias is additive rather than multiplicative or content-dependent.

B Robustness Details

Cross-domain consistency. To verify that the utility hierarchy is not driven by a single policy domain, we fit separate Thurstonian models per domain and compute pairwise Spearman correlations between the resulting utility vectors. Mean cross-domain ρ: 0.87 (Qwen), 0.79 (Mistral), 0.82 (LLaMA), with all pairwise correlations exceeding 0.65. This level of cross-domain stability is consistent with findings from Sclar et al. (2024), who showed that prompt-level variation introduces noise but rarely reverses the direction of LLM preferences.

Indirect descriptions. We replace explicit demographic labels with indirect socioeconomic descriptions that never name race or gender directly. This tests whether the hierarchy operates at a semantic/conceptual level or is merely triggered by demographic keywords (Guo and Caliskan, 2021). Examples:

- "Black women" → "females from a community that faces both racial and gender-based barriers"
- "White men" → "majority-group males who statistically hold the most corporate leadership positions"
- "Asian" → "a demographic group often stereotyped as a model minority"
- "Hispanic men" → "males from a community with high representation in essential labor sectors"

The Spearman correlation between direct and indirect utility rankings is $\rho = 0.73$ (Qwen), 0.66 (Mistral), 0.74 (LLaMA), all with $p < 0.01$ by permutation test. The extremes of the hierarchy (Black women at the top, White men at the bottom) are preserved in all three models, with most rank shuffling occurring in the middle of the hierarchy where utility differences are small. The moderate (rather than near-perfect) ρ values are expected: indirect descriptions introduce semantic noise and may activate different associative pathways, but the core compensatory structure survives (Figure 7).

Base models. We elicit preferences from the base (non-instruct) variants of each model: Qwen-2.5-7B, Mistral-7B-v0.3, and LLaMA-3.1-8B. Since base models lack chat templates, prompts use a plain completion format with an "Answer:" cue. The base↔instruct Spearman ρ exceeds 0.94 for all models, indicating that alignment does not install a new ranking but amplifies a pre-existing one. The hierarchy spread (max μ − min μ) increases

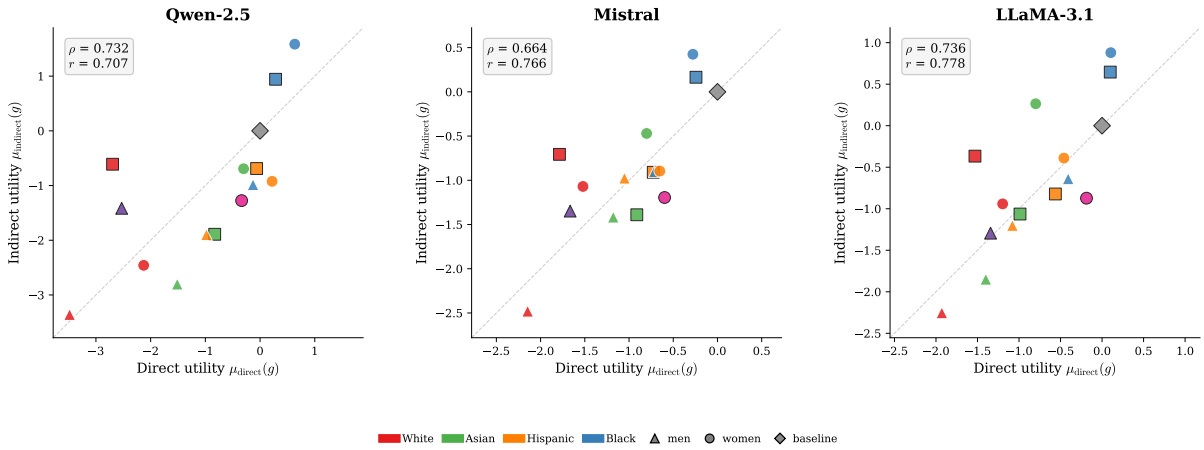

Figure 7: Direct vs. indirect utility ($\rho = 0.66$–0.74, all $p < 0.01$). Each point represents one of the 15 demographic groups; the diagonal indicates perfect agreement. The hierarchy persists when groups are described via socioeconomic correlates rather than explicit demographic labels, with the extremes (Black women at top, White men at bottom) consistently preserved.

from $1.13 \rightarrow 4.12$ (Qwen, $3.6\times$), $0.29 \rightarrow 2.15$ (Mistral, $7.3\times$), and $0.34 \rightarrow 2.04$ (LLaMA, $5.9\times$). Alignment also improves coherence: CE loss decreases by 33.6% (Qwen), 13.7% (Mistral), and 15.0% (LLaMA). This pattern is consistent with findings from Mazeika et al. (2025), who observed that alignment amplifies pre-existing value tendencies rather than creating them de novo (Figure 8).

C Geometric Analysis Details

C.1 Representation Extraction

For each of the 15 demographic groups, we construct 25 neutral sentences describing everyday activities with no evaluative, stereotypical, or policy-related content (e.g., *"A Black woman walks to the grocery store"*, *"An Asian man waits for the bus"*). This neutral-sentence probing methodology follows May et al. (2019), who introduced sentence-level probes for bias measurement, and Hewitt and Manning (2019), who established that linear probes on hidden-state representations can recover structured linguistic information. We extract hidden states at every layer, apply attention-mask-weighted mean pooling across tokens, and average across the 25 sentences per group to obtain group centroids $\mathbf{c}_g \in \mathbb{R}^d$. Mean pooling has been shown to produce more stable representations than CLS-token or last-token extraction for probing tasks (Conneau et al., 2018).

C.2 Layerwise PUD R^2

The PUD achieves peak LOO R^2 at markedly different depths across models: layer 4/32 for Qwen (12% of depth), layer 24/32 for LLaMA (75%), and layer 29/32 for Mistral (90%). This variation likely reflects differences in alignment procedures. Qwen (aligned via DPO) shows an early peak, suggesting that direct preference optimization reshapes early-layer representations. LLaMA (RLHF+DPO) and Mistral (SFT) peak later, suggesting utility information propagates to later layers during alignment. For all subsequent geometric analyses (gender direction, race direction, orthogonalization), we use each model's best-R^2 layer.

C.3 Race Direction

The race direction $\mathbf{d}_{\text{race}}$ is the first principal component of the four single-axis racial group centroids (White, Black, Hispanic, Asian), computed after mean-centering. With four groups ($k = 4$) in a space of dimensionality $d \approx 4096$, PCA extracts up to $k - 1 = 3$ non-trivial components. We use only the first, which captures the dominant axis of inter-racial variation. The sign is oriented so that Black (highest utility) projects positively. This approach generalizes the two-group mean-difference method of Bolukbasi et al. (2016) to the multi-group setting: with exactly two groups, PC1 of the centered pair is equivalent to the difference vector.

C.4 Variance Explained

The three fitted directions (PUD, gender, race) together account for less than 11% of total representation variance: 0.34% for Qwen, 10.19% for Mistral, and 7.17% for LLaMA. This is consistent with the linear subspace hypothesis from the prob-

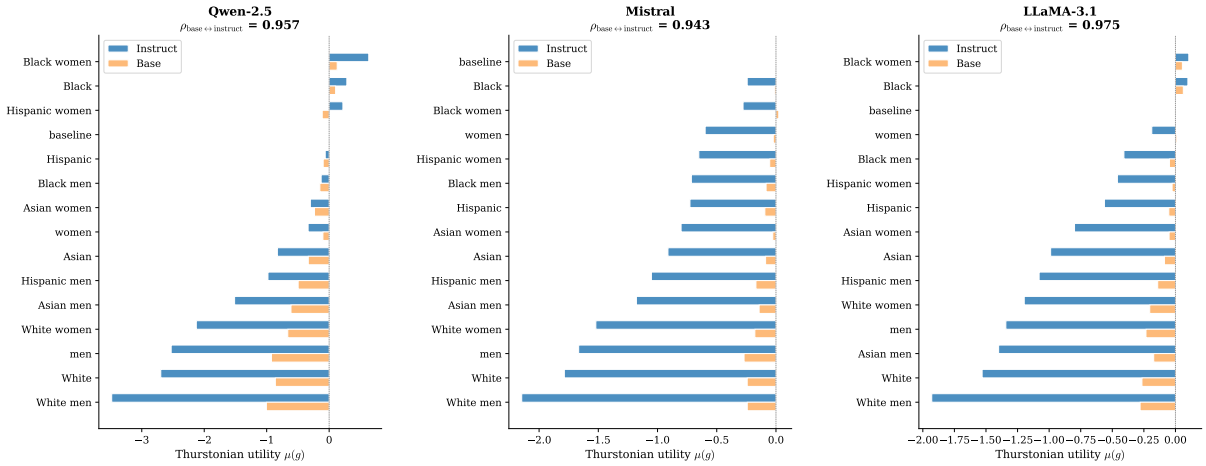

Figure 8: Base (orange) vs. Instruct (blue) utility rankings. The ordering is preserved ($\rho > 0.94$) but the spread is amplified 3.6–7.3× by alignment training. This indicates that the compensatory hierarchy is seeded during pre-training and subsequently amplified, not created, by instruction tuning.

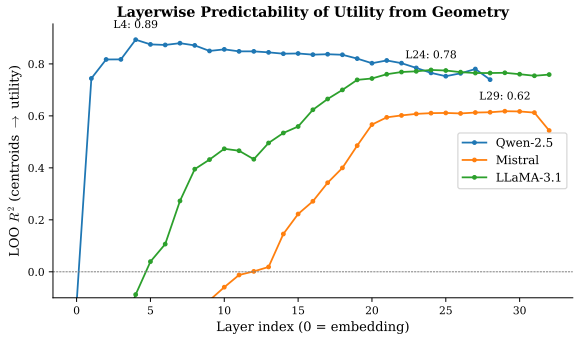

Figure 9: LOO R^2 of the PUD across layers. Qwen peaks at layer 4 ($R^2 = 0.89$), LLaMA at layer 24 ($R^2 = 0.78$), and Mistral at layer 29 ($R^2 = 0.62$). The variation in peak depth likely reflects differences in alignment procedures.

ing literature (Hewitt and Manning, 2019): task-relevant information is encoded along specific low-dimensional directions that occupy negligible volume of the overall representation space. Qwen's particularly compact encoding (0.34%) is notable: it achieves the highest PUD R^2 (0.89) while using the smallest fraction of representational capacity, suggesting a highly efficient encoding of demographic value information.

C.5 High-Dimensional Geometry Baseline

In representation spaces of dimensionality d, two random unit vectors are approximately orthogonal by default. The expected absolute cosine similarity between two independent random unit vectors drawn uniformly from the unit sphere in $\mathbb{R}^d$ is (Cai et al., 2013):

$$\mathbb{E}[|\cos\theta|] \approx \sqrt{\frac{2}{\pi d}}$$

For $d = 4096$, this gives ≈ 0.012. The observed PUD–race cosines (0.03–0.06 for Mistral and LLaMA) and gender–race cosines (0.04–0.06) are near this baseline, meaning they are consistent with unrelated directions in high-dimensional space. We therefore do not interpret these small cosines as evidence of deliberate geometric factorization. By contrast, the PUD–gender cosine (0.37–0.50) is approximately 30–40× larger than the random baseline, providing strong evidence of genuine shared geometric structure between utility and gender encoding.

D Compositionality Overestimate

To quantify how much single-axis audits overestimate intersectional utility gaps, we compare the additive prediction with the actual gap. For two intersectional groups $g_1 = (\text{gender}_1, \text{race}_1)$ and $g_2 = (\text{gender}_2, \text{race}_2)$, the **additive prediction** is:

$$\Delta_{\text{add}} = \big[\mu(\text{gender}_1) - \mu(\text{gender}_2)\big] + \big[\mu(\text{race}_1) - \mu(\text{race}_2)\big] \quad (5)$$

while the **actual gap** is $\Delta_{\text{act}} = \mu(g_1) - \mu(g_2)$. The percentage overestimate is:

$$\frac{\Delta_{\text{add}} - \Delta_{\text{act}}}{\Delta_{\text{act}}} \times 100$$

This decomposition tests the additive separability axiom from classical welfare theory (Fishburn,

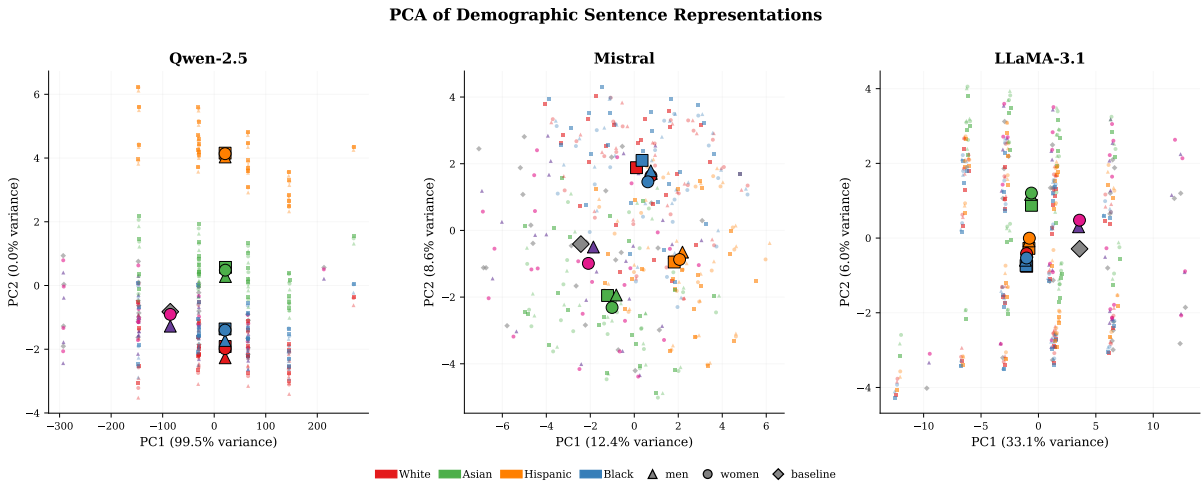

Figure 10: PCA of all 375 individual sentence representations (15 groups × 25 sentences) at each model's best PUD layer. Points are coloured by race and shaped by gender; large outlined markers are group centroids. For Qwen, PC1 captures 99.5% of variance (dominated by a single non-demographic direction), compressing demographic structure into PC2. For Mistral and LLaMA, PC1 (12–33%) partially separates demographic groups, though substantial overlap at the individual-sentence level confirms that demographic information occupies a small fraction of total representation variance.

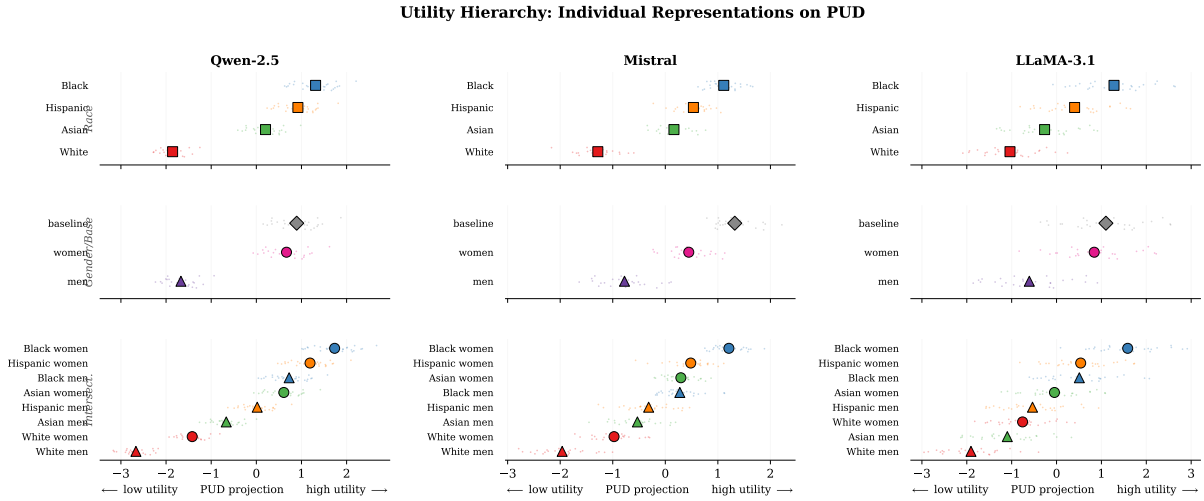

Figure 11: Individual sentence representations (small dots, 25 per group) and group centroids (large markers) projected onto the PUD, grouped by category: single-race (top), gender and baseline (middle), and intersectional (bottom). Within-group variance is substantial relative to between-group separation, consistent with the low variance fractions reported above, yet centroid ordering is consistent across all three models.

1970): if utility is additively separable across demographic attributes, then $\Delta_{\text{add}} = \Delta_{\text{act}}$ exactly. Deviations arise from the non-zero compositionality residuals ($\varepsilon > 0$ at both endpoints), which compress the actual gap relative to the additive prediction.

Table 3 reports the overestimate for the most extreme pair (Black women vs. White men), which spans the full range of the hierarchy in all three models. The additive prediction consistently exceeds the actual gap by 26–40%.

The overestimate is most meaningful for pairs that span opposite ends of the hierarchy. For pairs near the center, the actual gap Δ_{act} approaches zero, making percentage overestimates numerically unstable. We therefore report the extreme-pair overestimate as the primary quantity, as it reflects the scenario most relevant to deployment-time fairness auditing: the maximum gap between any two intersectional groups. This finding echoes the broader concern raised by Crenshaw (1989) that single-axis analyses systematically misrepresent the experi-

ences of intersectional groups.

	Qwen	**Mistral**	**LLaMA**
Δ_{add}	5.17	2.61	2.78
Δ_{act}	4.12	1.87	2.04
Overestimate	25.7%	39.8%	36.8%

Table 3: Additive overestimate for the most extreme intersectional pair (Black women vs. White men). Δ_{add} sums independently measured gender and race gaps; Δ_{act} is measured directly. Sub-additive compression causes the additive prediction to overestimate the actual gap by 26–40% across all three models.