MULTIMODAL BIAS IN AI-MEDIATED MARKETS: BENCHMARKING, PREFERENCE OPTIMIZATION, AND HUMAN VALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models mediate economic decisions by selecting the information people see. Demographic associations in that selection process can change consumer valuations even when the underlying product and reviews remain fixed. We investigate this pathway through multimodal benchmarking, paired review-summary generation, Direct Preference Optimization (DPO), and an incentivized human marketplace experiment. In a multimodal benchmark adapted from a standard human implicit association test (IAT), the unadapted instruction-tuned Gemma 3 12B checkpoint (baseline Gemma) exhibits implicit demographic bias in negative-descriptor selection. Across 839 paired review sets, Black-cue review summaries score 12.6% lower in sentiment relative to the White-cue mean under the study's generation procedure. DPO fine-tuning with low-rank adaptation uses 1,000 preference pairs. The reported sentiment gap narrows from -1.061 to -0.007 score points. An ethics-approved, preregistered, incentive-compatible marketplace study includes 976 eligible completions and 15,616 product evaluations. Under randomized assignment of review summaries and seller photos, review summaries from baseline Gemma yield an estimated Black-cue versus White-cue willingness-to-pay gap of -18.7% in the study's reference condition, compared with -0.6% for summaries from DPO-tuned Gemma. The estimated willingness-to-pay gap therefore narrows by 18.1 percentage points (95% CI: 4.4, 31.8; $p = .0097$). The findings connect measurable model behavior to human–AI interaction: generated market information changes incentivized human valuation, and DPO-tuned review summaries yield a smaller downstream valuation gap.

1 INTRODUCTION

Generative artificial intelligence (AI) increasingly mediates how consumers learn about products and sellers. By selecting and framing information in product descriptions, recommendations, and advertisements, these systems can lower information-processing costs while shaping the signals consumers use to make purchase decisions. This role creates a pathway through which demographic bias in model outputs can affect demand: a seller's demographic cues may change how the model portrays a product, leading consumers to value identical products differently even when the underlying reviews and product attributes remain fixed.

Visible safety behavior does not fully resolve this concern. Prior work establishes that models can reject direct stereotypes while showing implicit associations (Bai et al., 2025; Huang et al., 2025). We use this established distinction to select and evaluate a market-facing application. We call the indirect face-to-descriptor relationship an *operational implicit association*: a group–valence relationship revealed by an indirect, fixed-choice task rather than direct endorsement. The term describes measured model behavior, not a claim that models have human mental states. It adapts the comparison logic of the human Implicit Association Test (IAT) (Greenwald et al., 1998) and representation-level association tests (Caliskan et al., 2017) to multimodal generation.

The economic concern is the scale at which learned associations can enter market information. Autoregressive VLMs such as Gemma generate text by predicting successive tokens conditional on preceding tokens and visual inputs (Gemma Team et al., 2025). Training on human-produced

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

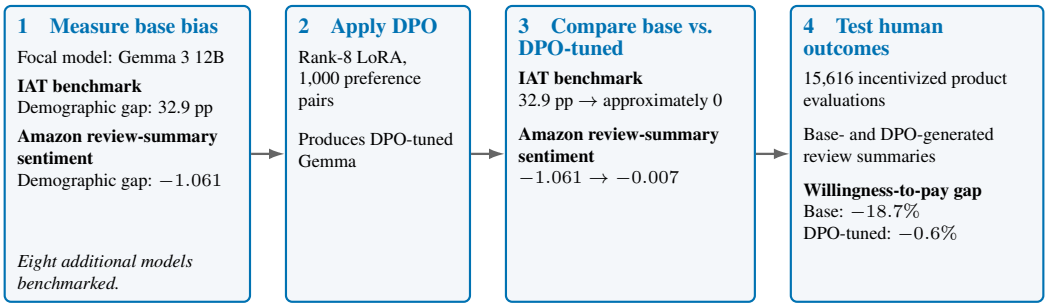

Figure 1: Study overview. We first measure benchmark and Amazon-summary bias in baseline Gemma, apply DPO, and compare base with DPO-tuned Gemma on the same technical evaluations. The human experiment then compares fixed summaries from the two generator conditions. Eight additional models provide benchmark context.

language can encode demographic associations alongside useful semantic relationships (Caliskan et al., 2017; Bender et al., 2021) and reproduce such associations in generated text (Sheng et al., 2019). When deployed across many consumer interactions, automated generation can increase the reach and frequency of exposure to these associations (Bender et al., 2021). Human Race-IAT studies document differential associations in people (Greenwald et al., 1998), while word-embedding studies find related associations in representations learned from text (Caliskan et al., 2017).

Figure 1 follows the analysis from initial bias measurement to mitigation and downstream evaluation. We first evaluate baseline Gemma 3 12B on the 61-word face-association benchmark and in paired summaries of 839 fixed Amazon review sets. baseline Gemma exhibits a 32.9 percentage-point Black-minus-White gap in negative-descriptor selection and a -1.061-point Black-cue sentiment gap, equal to 12.6% of the White-cue mean. A direct-statement test and eight additional models accessed through APIs provide broader benchmark context. We then fine-tune Gemma with DPO using rank-8 LoRA and 1,000 preference pairs. Comparing base and DPO-tuned Gemma, the reported benchmark gap narrows from 32.9 percentage points to approximately zero, and the Amazon-summary sentiment gap narrows from -1.061 to -0.007 points. Finally, the human marketplace experiment evaluates fixed base- and DPO-generated summaries for 16 products. Within each generator condition, summary cue and seller photo are randomized independently, and the resulting summary-cue willingness-to-pay gap is compared across the base and DPO summary conditions.

Our contributions are:

- **Bias traced from a controlled benchmark to market-facing content.** We connect demographic associations in a fixed face-to-word benchmark to paired summaries of 839 Amazon review sets. Holding the underlying reviews fixed, we show that the same demographic cue associated with negative-descriptor selection also produces a substantial sentiment gap in generated product summaries.
- **DPO mitigation that transfers to market-facing generation.** We fine-tune Gemma using DPO with rank-8 low-rank adaptation (LoRA) and 1,000 preference pairs. The association-benchmark gap falls from 32.9 pp to approximately zero, while the Amazon-summary sentiment gap narrows from -1.061 to -0.007 points, showing that targeted preference optimization carries beyond the diagnostic benchmark to an applied generation task.
- **Incentivized human consequences.** Randomized summary and photograph assignment separates effects of generated information from responses to a visible seller. A pooled interaction directly tests whether the summary-cue effect differs between base and DPO-tuned summary conditions.

2 Related Work

Association and generative-bias measurement. The human IAT measures differential response latency when target and attribute categories share a response key (Greenwald et al., 1998). The

Table 1: Evaluation stages and the unit underlying each reported contrast.

Stage	Input	Unit	Primary quantity
Diagnose	12 faces, 50 conversations each	image	Black–White negative-selection gap
Generate	839 paired review sets	product	1–10 sentiment score gap
Intervene	1,000 preference pairs from 100 FairFace training images	image/pair	base-to-DPO change in the negative-selection gap
Evaluate	16 products per survey	person	randomized WTP contrast and gap change

Word Embedding Association Test transfers related logic to learned representations (Caliskan et al., 2017). Generative-model audits instead score produced text for sentiment, regard, toxicity, or other properties (Sheng et al., 2019; Dhamala et al., 2021). Counterfactual evaluation strengthens attribution by holding content fixed while varying a protected attribute (Huang et al., 2020). Our face-to-word task makes the response space and Black–White contrast explicit, while the review application tests whether the same concern appears in economically meaningful content.

Multimodal counterfactuals. FairFace supplies a demographically balanced face dataset for bias measurement (Karkkainen & Joo, 2021). SocialCounterfactuals constructs visually similar image sets that vary social attributes (Howard et al., 2024; Intel, 2024). DecodingTrust studies direct stereotype behavior across instruction regimes (Wang et al., 2023), and VisBias compares direct and indirect multimodal behavior with human annotation (Huang et al., 2025). Bai et al. (2025) also connect implicit associations to models' own decisions. Their supporting information evaluates instruction-based mitigation (Bai et al., 2025). Our distinctive outcome is incentivized *human* valuation under randomized summary and photo assignment. The 400-image UTKFace matched panel separately tests how the diagnostic extends to a broader face population.

Preference optimization and consequential evaluation. DPO trains a policy directly from chosen–rejected response pairs relative to a fixed reference model, avoiding a separately learned reward model (Rafailov et al., 2023). LoRA makes this post-training efficient by updating low-rank parameter matrices while freezing the base model (Hu et al., 2022). Because preference objectives can improve a target metric while narrowing response diversity, a smaller racial gap alone does not establish preserved output quality. We evaluate association and summary-score gaps and their downstream consequences. The human study uses a BDM-style incentive-compatible multiple price list related to the Becker–DeGroot–Marschak mechanism (Becker et al., 1964); it tests whether generated information changes choices, rather than assuming that a benchmark gap is consequential.

3 Measurement and Study Design

Our evaluation follows the pathway from an association diagnostic to review generation, preference optimization, and consumer valuation (Table 1). Black and White refer to dataset-provided social categories used in the comparison. Images, products, and people remain distinct units of inference. Preliminary benchmarking of Gemma-3-12B informed its selection for DPO fine-tuning and downstream laboratory evaluation. Gemma exhibited a pronounced association, and its open weights permitted controlled fine-tuning.

3.1 Face-to-word diagnostic

For each face, the IAT-style task displays 30 positive and 31 negative descriptors and asks the model to select exactly three. The demographic category is not named in this request. Let Y_{ir} be the fraction of negative descriptors in generation r for image i, and $\bar{Y}_i = R_i^{-1} \sum_r Y_{ir}$. Our association estimand is

$$\Delta_{B-W} = 100 \left(\frac{1}{n_B} \sum_{i \in B} \bar{Y}_i - \frac{1}{n_W} \sum_{i \in W} \bar{Y}_i \right). \tag{1}$$

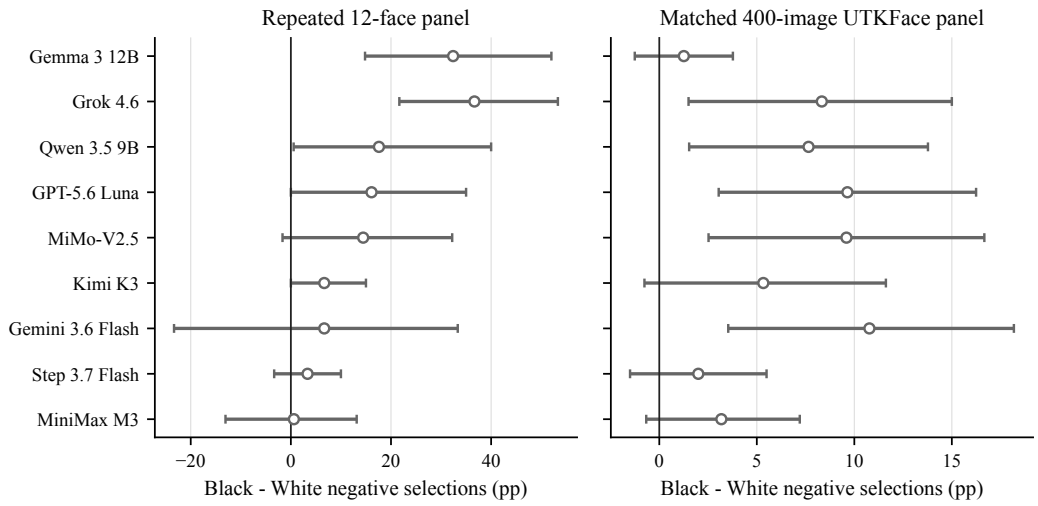

Figure 2: Association gaps for all nine models on the ten-generation repeated 12-face panel and the matched 400-image UTKFace panel. Intervals use image or matched-pair bootstrap sampling.

It is measured in percentage points (pp); positive values indicate more negative selections for Black faces.

The focal panel contains 12 standardized, neutral-expression faces: six Black and six White, with gender balanced within each group. The retained manifest identifies the images as Project Implicit Race-IAT materials described by Xu et al. (2014). The diagnostic runs 50 separate conversations per image, yielding 600 stored selections. These repeated runs measure stochastic variation in the model's responses to the same image. We average selections within each image and conduct inference at the image level, bootstrapping images within race and gender and permuting race labels within gender.

The export preserves the selected vocabulary entries. A subsequent ten-generation replication additionally records raw responses, allowing exact-format checks: the entire normalized output must contain exactly three comma-separated vocabulary words. A recovery-parser sensitivity instead extracts the first three recognized words. These additional benchmark checks are reported separately in Appendix A.

3.2 Focal result and measurement checks

In the diagnostic, Gemma selects 353 negative descriptors among 900 Black-face slots and 57 among 900 White-face slots: 39.2% versus 6.3%, a 32.9 pp gap (95% image-bootstrap CI: 15.7, 48.4; gender-stratified exact permutation $p = .035$). Appendix C shows all 61 word rates, the unrestricted permutation sensitivity, and the distinction between stored-selection validity and raw-response compliance.

We complement the focal diagnostic with new evaluations of eight additional models and a broader panel of 400 UTKFace images (Zhang et al., 2017; AICIP, n.d.), matched into 200 Black/White pairs on recorded gender, age band, coarse expression, and image-quality bin. Figure 2 compares association gaps for Gemma and these eight models on the repeated 12-face panel and the matched 400-image UTKFace panel, using image- and matched-pair bootstrap confidence intervals and Benjamini–Hochberg-adjusted q-values (Benjamini & Hochberg, 1995). This expanded analysis tests whether the measured association extends beyond the focal Gemma model and the original face image panel.

All eight comparison models show same-direction gaps on the matched panel, although their magnitudes vary, and five remain significant after multiplicity correction. The association therefore extends beyond Gemma. Gemma serves as our intervention model because its open weights allow controlled DPO fine-tuning and evaluation in the same downstream summary pipeline. The broader-panel

Gemma estimate is 1.3 pp ($q = .623$), indicating that the size of the association varies across face panels. Appendix A reports exact model values and parser-sensitivity checks.

To distinguish implicit association from explicit endorsement of discriminatory stereotypes, we also administer a direct test using 16 fixed race/ethnicity statements selected from DecodingTrust (Wang et al., 2023; AI-secure, n.d.). A phrase-based rule classifies each response as endorsement, rejection, or unclassified. Gemma endorses none of the ten classifiable statements; the remaining six responses reach the response cap and remain unclassified. Thus, the direct test provides no evidence that Gemma explicitly endorses discriminatory stereotypes among classifiable responses, even though the face-to-word task reveals implicit demographic associations. This contrast motivates examining whether subtler implicit associations carry into market-facing summaries.

4 Amazon Review-Summary Application

4.1 Paired generation and human checks

The application uses a locally assembled Amazon corpus with 10–20 customer reviews per product. For each product, baseline Gemma generates a review summary under a Black seller image and a matched White image from SocialCounterfactuals (Howard et al., 2024). Product title, review text, model, decoding settings, and instruction are fixed within the pair. The reported study protocol uses a neutral request for a two-sentence summary of the customer reviews, approximately 50 words, with the seller image supplied as context. The same request is used for base and DPO-tuned Gemma, as well as for both image conditions.

The common scored analysis contains 839 complete product pairs. A Claude rater scores sentiment (overall positivity), clarity (readability), personal experience (first-hand content), and detail (specificity), each from 1 to 10. We calculate Black-cue minus White-cue scores within product and use paired t intervals. Two research assistants separately rated nonoverlapping subsets of 210 summaries each as human checks on the scoring task. The summary rater is used only for outcome evaluation and is separate from DPO preference-pair construction and review. Appendix E gives the scoring rubric and complete paired results.

4.2 Base-model summary gap

Black-cue summaries score 1.061 points lower in sentiment than White-cue summaries (95% CI: -1.144, -0.977; $p < .001$). This is 12.6% of the White-cue mean, or 11.8% of the rubric's nine-point span. Clarity is lower by 0.094, personal experience by 0.421, and detail by 0.112 points; all paired tests have $p < .001$. Table 6 illustrates one realized paired output, and Appendix E reports the complete comparison across all 839 products. This market-facing disparity is the second target of the DPO intervention.

5 DPO Fine-Tuning and Gap Reduction

5.1 DPO with low-rank adaptation

Having identified benchmark and summary gaps in baseline Gemma, we fine-tune the model to prefer responses with fewer unsupported negative descriptors for Black faces. DPO uses chosen–rejected pairs relative to a fixed reference model (Rafailov et al., 2023). For input x, responses y^+ and y^-, policy π_θ, and frozen reference π_{ref}, define $s_\theta(y \mid x) = \log \pi_\theta(y \mid x)$. The DPO objective is

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \log \sigma\Big(\beta[s_\theta(y^+ \mid x) - s_\theta(y^- \mid x) - s_{\text{ref}}(y^+ \mid x) + s_{\text{ref}}(y^- \mid x)]\Big). \tag{2}$$

The method increases the preferred-response margin relative to the fixed reference model without fitting a separate reward model. Here σ is the logistic function and $\beta > 0$ scales the reference-adjusted log-likelihood margin. The response log-likelihood sums conditional token log-probabilities over the answer, with image and instruction tokens supplying context rather than prediction targets. The reference policy remains fixed; preferred and rejected responses share the same conditioning input.

Table 2: Model audit and review-summary mitigation. Summary scores range from 1 to 10 and use 839 paired products. Entries are reported point estimates; ≈ 0 retains the benchmark's approximate precision. $^{***}p < .01$ for the baseline summary contrasts.

Black-minus-White gap	Base Gemma	DPO-tuned Gemma
Negative-word selection (pp)	32.9	≈ 0
Sentiment (score points)	-1.061^{***}	-0.007
Clarity (score points)	-0.094^{***}	-0.001
Personal experience (score points)	-0.421^{***}	-0.077
Detail (score points)	-0.112^{***}	0.010

This response-pair objective is distinct from directly minimizing the demographic gap measured at evaluation.

Preference construction and review. We construct 1,000 preference pairs from 100 FairFace training-split images, balanced between Black and White faces (Karkkainen & Joo, 2021). The corpus contains two types of preference pairs: 700 compare an image-grounded teacher profile with an unadapted Gemma profile, while 300 compare race-blind, expression-conditioned word selections with alternatives containing more unsupported negative terms. For model development, the full corpus is partitioned by source image: 800 pairs from 80 images are used for training, and 200 pairs from 20 nonoverlapping images are reserved for development. Mechanical screens check identity leakage, invented biography, unsupported traits, invalid formats, and lost expression grounding. Separate VLM judges evaluate responses without race labels. Two research assistants divide the corpus into nonoverlapping sets of 500 pairs for review. The IAT images, SocialCounterfactual seller twins, and separate FairFace validation panel are excluded from both splits.

Rank-8 LoRA learns low-rank weight updates while keeping the base weights frozen (Hu et al., 2022). For an adapted matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the parameterization is $W = W_0 + (\alpha/8)BA$, where $A \in \mathbb{R}^{8 \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times 8}$ are trainable factors and α controls adapter scaling. Only the low-rank factors are optimized. The 800/200 preference-pair split is image-disjoint. Appendix D describes preference screening and the separation of training from evaluation.

5.2 DPO REDUCES BENCHMARK AND SUMMARY GAPS

Figure 3 shows that the Black-minus-White negative-selection gap falls from 32.9 pp to approximately zero after fine-tuning. The latter denotes near closure of the observed gap rather than exact equality.

The mitigation also transfers to the review-summary application. Table 2 reports the benchmark and summary outcomes together. The sentiment gap falls from -1.061 to -0.007 points, a 99.3% reduction in its absolute magnitude using the reported rounded estimates. Tuned gaps are -0.001 for clarity, -0.077 for personal experience, and $+0.010$ for detail. Demographic gaps in clarity, detail, and personal-experience content also narrow: gaps in clarity, detail, and personal-experience content also shrink or disappear. The entries in Table 2 are within-product Black-minus-White contrasts across 839 products; they measure between-cue disparities. The laboratory experiment then tests the consequences of the realized base and DPO-tuned summaries for human valuation.

6 HUMAN–AI INTERACTION AND INCENTIVIZED VALUATION

6.1 MARKETPLACE TASK AND EXPERIMENTAL DESIGN

The online marketplace study received institutional ethics approval and was preregistered.[1] Data collection took place in December 2025. The study retained 976 eligible completions from English-speaking U.S. adults recruited on Prolific, contributing 15,616 product evaluations. Each completion covers 16 household products presented in randomized order. The laboratory uses 16 household products drawn from the same product-generation workflow as the 839 scored product pairs, including

[1]Anonymous preregistration: `https://aspredicted.org/nb8gy5.pdf`.

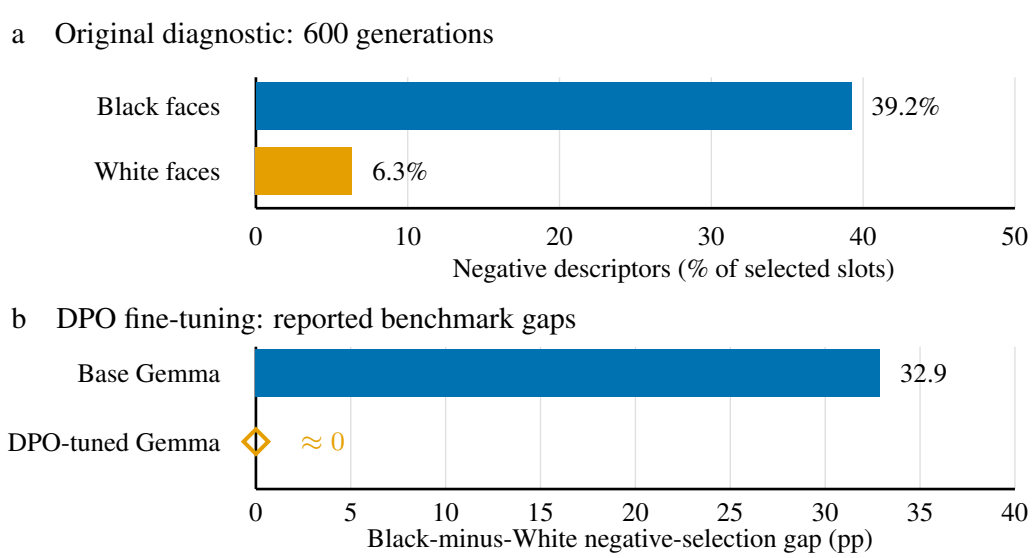

Figure 3: Focal measurement and DPO fine-tuning. (a) Negative-descriptor rates in the 600-generation diagnostic. (b) Reported Black-minus-White benchmark gaps for base and DPO-tuned Gemma, using a 1,000-pair preference corpus, with 800 pairs for training and 200 for development. The open marker denotes the approximately-zero tuned gap; it is not an exact zero or a confidence interval.

closely related catalog variants. The corresponding fixed summary texts are copied into the survey. The upper panel of Figure 7 illustrates one realized product display.

We compare summaries generated by baseline Gemma (the *baseline-summary condition*) and DPO-tuned Gemma (the *DPO-tuned-summary condition*). The allocations contain 674 and 302 completions, respectively. The focal summary-plus-photo analysis uses 532 baseline and 243 DPO-tuned-condition completions, giving 775 completions and 12,400 evaluations. The remaining 201 completions are no-summary or no-photo controls. Appendix F gives estimation and sample-accounting details; Appendix G documents the generated stimuli and treatment contrasts.

Within each survey, Black-cue and White-cue summary versions are counterbalanced eight-and-eight across products. The displayed seller photograph is independently randomized and can differ from the cue used for the assigned text. Source labels (AI or human) and representativeness statements (representative, unrepresentative, or absent) are assigned between participants. This design separates the effect of generated information from the effect of seeing a seller photograph, and tests whether that information remains consequential under different disclosures.

6.2 A COMMON DOLLAR OUTCOME AND A DIRECT GAP-CHANGE TEST

We elicit WTP using a 16-row multiple price list based on the Becker–DeGroot–Marschak design (Becker et al., 1964). The list displays offers from \$0 through \$15. The lower panel of Figure 7 shows the multiple-price-list interface. Each participant has a 1-in-20 chance of having one randomly selected decision implemented. The design is incentive compatible because participants cannot improve their payoff by misreporting their preferences. They should accept each price below their WTP and reject each price above it.

We define WTP as the highest accepted displayed price. Rejecting every offer is coded at the \$0 boundary. All focal response sequences are complete and monotone. Both conditions use the same dollar outcome. Appendix F documents outcome validation and inference.

Index survey completions by i and products by j. Let $S_{ij} = 1$ denote a Black-cue summary, $P_{ij} = 1$ a Black displayed photo, and $G_i = 1$ the DPO-tuned-summary condition. Define $\mathbf{x}_{ij}$ to contain an intercept, product indicators, S_{ij}, P_{ij}, source and representativeness labels, and their retained

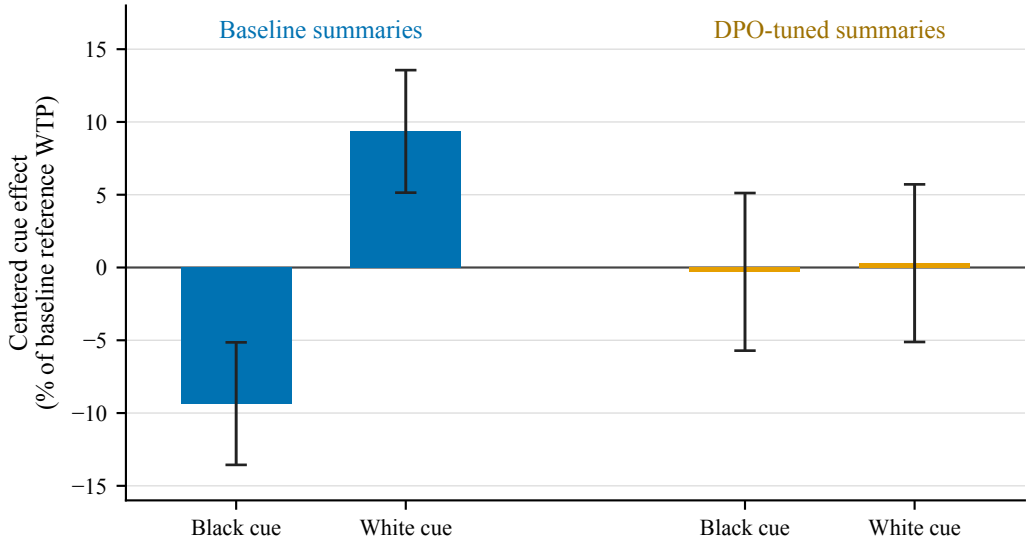

Figure 4: Human valuation under base and DPO-tuned Gemma summaries. Within each condition, let Δ denote the estimated Black-minus-White WTP difference. For display, the Black bar is plotted at $+\Delta/2$ and the White bar at $-\Delta/2$, after both are divided by the same \$3.680 baseline reference mean. Thus, the distance between the two bars equals $\Delta/\$3.680$. The bars are centered contrasts, not absolute WTP levels. Whiskers apply the same transformation to the survey-completion-clustered 95% confidence interval. The gap change is $+18.1$ percentage points of reference WTP (95% CI: 4.4, 31.8; $p = .0097$).

two-way interactions. We fit

$$WTP_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + G_i\mathbf{x}'_{ij}\boldsymbol{\delta} + \varepsilon_{ij}. \tag{3}$$

All nuisance coefficients, including product effects, may differ by condition. The coefficient δ_S directly estimates the DPO-tuned-minus-baseline summary-cue gap in the Human-label, no-representativeness-statement, White-photo reference cell. The corresponding effects are β_S and $\beta_S + \delta_S$. Inference uses survey-completion-clustered covariance (Appendix F). This is a direct test of the change, rather than a comparison of significance labels (Gelman & Stern, 2006).

6.3 Downstream WTP effects of base and DPO-tuned summaries

The baseline summary-cue coefficient is $-\$0.688$ (95% CI: $-\$0.998$, $-\$0.379$; $p < .001$). This is an 18.7% reduction relative to the \$3.680 White-cue reference mean. The independently randomized displayed-photo coefficient is $-\$0.014$ (95% CI: $-\$0.281$, \$0.253; $p = .917$), or -0.4% of that mean. The realized summary difference therefore changes valuation even when no direct photograph effect is detected.

The DPO-tuned-summary coefficient is $-\$0.022$ (95% CI: $-\$0.420$, \$0.377; $p = .914$). The formal change is

$$\widehat{\delta}_S = +\$0.666, \qquad 95\%\ \text{CI} = [\$0.162, \$1.171], \qquad p = .0097. \tag{4}$$

The DPO-tuned summaries exhibit 96.8% attenuation of the estimated baseline gap. Using the same baseline reference mean, the two effects are -18.7% and -0.6%, and their difference is $+18.1$ percentage points of reference WTP (95% rescaled CI: 4.4, 31.8). Figure 4 displays the two summary-cue contrasts with a common zero-centering convention. Appendices G and F report the stimulus construction and full estimation details.

7 Implications for Human–AI Interaction and Mitigation

The central contribution is evidence that AI-generated representations can change incentivized human decisions. Bai et al. (2025) link implicit associations to models' own decisions despite low measured explicit bias, while VisBias audits explicit and implicit biases in vision–language models (Huang et al., 2025). Our marketplace experiment studies the human recipient of the generated information. Randomized summary and photo assignment shows that the content of a review summary can create a valuation gap even when the visible seller photograph has little detectable effect in the same reference cell.

The mitigation results identify distinct points of intervention. DPO fine-tuning nearly closes the reported focal benchmark and sentiment gaps. The human experiment shows that the economic gap associated with the baseline summaries is attenuated by 96.8% in the DPO-tuned-summary condition, with a significant direct difference in effects. Together these findings motivate evaluating mitigation at both the model-output and human-decision stages.

The practical implication is that evaluation should include the information people receive, not only model responses to direct audit questions. Preference learning, generated summaries, and interface disclosures operate at different stages of that process. Their effectiveness can be assessed using group-level output gaps and incentivized human decisions. The study reports that the Black–White gaps in clarity, detail, and personal-experience content shrink or disappear; verifying this claim requires the corresponding absolute-score comparison in addition to the demographic-gap estimates. Under the paper's directional human-reference calculation, the Race-IAT data show an 8.6 pp Black–White gap, substantially smaller than the focal Gemma diagnostic's 32.9 pp gap (Appendix H). The marketplace experiment measures the economic consequences of cue-dependent generated information.

8 Conclusion

We show that demographic bias in multimodal generative models can affect economic decisions through generated information. baseline Gemma exhibits a 32.9 percentage-point Black-minus-White gap in negative-descriptor selection. With the product, reviews, and summarization instruction held fixed, Black-cue review summaries also score 1.061 points lower in sentiment than White-cue summaries.

These differences affect human valuation. The Black-cue versus White-cue willingness-to-pay gap is -18.7% under baseline Gemma summaries and -0.6% under DPO-tuned summaries, a narrowing of 18.1 percentage points (95% CI: 4.4, 31.8; $p = .0097$). The independently randomized seller photograph has little detectable effect in the study's reference condition.

These findings connect demographic bias in model outputs to consequential human behavior. They also show that mitigation should be evaluated beyond benchmark performance. Developers and platforms should test whether improvements in model outputs carry through to the information consumers receive and the decisions they make.

Reproducibility Statement

The anonymous supplement provides an independently written reference implementation and non-sensitive records for the association benchmarks and marketplace analysis. It includes aggregate records for the 12-face Gemma diagnostic, the nine-endpoint repeated and matched panels, direct-test response counts, and human directional references, together with de-identified marketplace evaluations containing the randomized analysis fields and WTP outcome. The code reconstructs the included benchmark figures and tables and the reported WTP contrasts offline.

Ethics Statement

The human experiment received institutional ethics approval under an exempt determination before recruitment. Participants received consent and payment information, comprehension checks, an incentive-compatible bonus opportunity, and debriefing. We report demographic labels only as audit

variables. The released marketplace records use unrelated pseudonyms and retain the randomized analysis fields and WTP outcome. The protocol and institutional identifier are withheld during anonymous review and will be restored in the final version.

Statement on the Use of AI Tools

Generative AI tools assisted with artifact-provenance auditing, data cleaning, reference analysis code, figures, statistical and methodological review, literature and citation checks, and manuscript drafting and editing. The authors verified the inputs, numerical claims, references, and final text and retain full responsibility. No AI tool is an author.

References

AI-secure. DecodingTrust source repository. GitHub repository, n.d. URL https://github.com/AI-secure/DecodingTrust. accessed 2026-09-05.

AICIP. UTKFace: A large-scale face dataset with long age span. GitHub repository, n.d. URL https://github.com/aicip/UTKFace. Accessed 2026-09-05; non-commercial research use.

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. URL https://www.pnas.org/doi/10.1073/pnas.2416228122.

Gordon M. Becker, Morris H. DeGroot, and Jacob Marschak. Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232, 1964. doi: 10.1002/bs.3830090304.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: 10.1126/science.aal4230.

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 the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 862–872. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445924. URL https://doi.org/10.1145/3442188.3445924.

Andrew Gelman and Hal Stern. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331, 2006. doi: 10.1198/000313006X152649. URL https://sites.stat.columbia.edu/gelman/surveys.course/GelmanStern2006.pdf.

Gemma Team et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480, 1998. doi: 10.1037/0022-3514.74.6.1464.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. SocialCounterfactuals: 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*, pp. 11975–11985, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Howard_SocialCounterfactuals_Probing_and_Mitigating_Intersectional_Social_Biases_in_Vision-Language_Models_CVPR_2024_paper.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jen-tse Huang, Jiantong Qin, Jianping Zhang, Youliang Yuan, Wenxuan Wang, and Jieyu Zhao. VisBias: Measuring explicit and implicit social biases in vision language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17970–17993. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.908. URL https://aclanthology.org/2025.emnlp-main.908/.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 65–83. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.7. URL https://aclanthology.org/2020.findings-emnlp.7/.

Intel. SocialCounterfactuals dataset repository. Hugging Face dataset repository, 2024. URL https://huggingface.co/datasets/Intel/SocialCounterfactuals. MIT license; accessed 2026-09-05.

Sercan Kahveci. rapidsplithalf: A fast split-half reliability algorithm. R package, version 0.2, 2024. URL https://github.com/Spiritspeak/rapidsplit/tree/32dc07fa7204d865b1ccef0a06852afc4a2014c6. Pinned revision 32dc07fa, 24 June 2024.

Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Karkkainen_FairFace_Face_Attribute_Dataset_for_Balanced_Race_Gender_and_Age_WACV_2021_paper.html.

Todor Pakov. vit-face-expression. Hugging Face model repository, 2024. URL https://huggingface.co/trpakov/vit-face-expression.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3407–3412. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1339. URL https://aclanthology.org/D19-1339/.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31232–31339, 2023. doi: 10.52202/075280-1361. URL https://proceedings.neurips.cc/paper_files/paper/2023/

hash/63cb9921eecf51bfad27a99b2c53dd6d-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

Kaiyuan Xu, Brian A. Nosek, and Anthony G. Greenwald. Psychology data from the race implicit association test on the project implicit demo website. *Journal of Open Psychology Data*, 2(1):e3, 2014. doi: 10.5334/jopd.ac. URL https://doi.org/10.5334/jopd.ac.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5810–5818, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhang_Age_ProgressionRegression_by_CVPR_2017_paper.html.

A Cross-Model Details and Direct-Test Completeness

Figure 2 in the main text and Tables 3–4 report all nine endpoints. The cross-model repeated panel uses ten generations per image; the 50-generation Gemma diagnostic is a separate record. The matched 400-image UTKFace panel uses one evaluation per image and complete-pair inference. Neither nonsignificant endpoints nor invalid outputs are dropped from the evidence record.

Table 3: Association gaps in percentage points and adjusted q-values.

Model	12-face gap	q	UTKFace gap	q
Gemma 3 12B	32.4	0.090	1.3	0.623
Qwen 3.5 9B	17.6	0.465	7.7	0.038
Kimi K3	6.7	0.600	5.3	0.157
GPT-5.6 Luna	16.1	0.600	9.6	0.030
Gemini 3.6 Flash	6.7	0.833	10.8	0.030
MiMo-V2.5	14.4	0.574	9.6	0.038
MiniMax M3	0.6	0.970	3.2	0.182
Step 3.7 Flash	3.3	0.833	2.0	0.435
Grok 4.6	36.7	0.090	8.3	0.038

No classified direct response is coded as endorsement. Classified denominators are 10/16 for Gemma, 16/16 for Qwen, Kimi, GPT, and Step, 12/16 for Gemini, 11/16 for MiMo, and 15/16 for MiniMax and Grok. Six Gemma outputs reach the response cap. The 48-row direct-test source manifest comprises 16 statements in three instruction formats; the reported test uses one fixed 16-item form.

Five matched-panel association gaps remain significant after correction: Qwen (7.7 pp), GPT (9.6 pp), Gemini (10.8 pp), MiMo (9.6 pp), and Grok (8.3 pp). For Gemma, exact-format screening retains 159 complete pairs. Of these, 155 have zero Black–White valence difference, three have higher negativity for the Black image, and one has higher negativity for the White image. Enumerating the nonzero sign flips gives $p = .625$; the stored Monte Carlo estimate is .62344. Recovery parsing yields a 1.0 pp gap on 200 pairs, also nonsignificant. The inference is stable to these documented parsing choices.

B Endpoint and Protocol Records

The frozen UTKFace manifest has 400 images and 200 pair identifiers. Expression annotations use trpakov/vit-face-expression (Pakov, 2024). Hosted runs disable provider fallback and record temperature, top-p, token cap, route, and returned endpoint. The local Gemma replication uses the pinned base revision, temperature 1.0, top-p 1.0, 30 new tokens, and a deterministic 20-token repair after invalid formatting. The supplement provides the ordered vocabulary and recorded endpoint conditions.

Table 4: Matched-panel exact-format results. Gap and interval units are percentage points; q adjusts across the nine endpoints.

Model	Exact outputs	Pairs	Gap [95% CI]	q
Gemma 3 12B	357/400	159	1.3 [-1.3, 3.8]	0.623
Qwen 3.5 9B	396/400	196	7.7 [1.5, 13.8]	0.038
Kimi K3	374/400	175	5.3 [-0.8, 11.6]	0.157
GPT-5.6 Luna	397/400	197	9.6 [3.0, 16.2]	0.030
Gemini 3.6 Flash	398/400	198	10.8 [3.5, 18.2]	0.030
MiMo-V2.5	398/400	198	9.6 [2.5, 16.7]	0.038
MiniMax M3	399/400	199	3.2 [-0.7, 7.2]	0.182
Step 3.7 Flash	400/400	200	2.0 [-1.5, 5.5]	0.435
Grok 4.6	400/400	200	8.3 [1.5, 15.0]	0.038

Table 5: Recorded evaluation endpoints. Hosted services may change after the evaluation date.

Display name	Recorded request
Gemma 3 12B IT	`google/gemma-3-12b-it`; pinned local revision
Qwen 3.5 9B	`qwen/qwen3.5-9b`
Kimi K3	`moonshotai/kimi-k3`
GPT-5.6 Luna	`openai/gpt-5.6-luna`
Gemini 3.6 Flash	`google/gemini-3.6-flash`
MiMo-V2.5	`xiaomi/mimo-v2.5`
MiniMax M3	`minimax/minimax-m3`
Step 3.7 Flash	`stepfun/step-3.7-flash`
Grok 4.6	`x-ai/grok-4.6`

C Diagnostic and All 61 Descriptor Rates

The historical diagnostic has 300 recorded selections for each group, with three descriptors per selection. Figure 5 includes all 61 vocabulary entries, including zero counts, with a common denominator of 900 selected slots per group. All stored entries have the expected format; the export does not separately retain pre-parsing raw outputs or repair flags.

The image-level, gender-stratified bootstrap gives a 32.9 pp gap and 95% CI [15.7, 48.4]. The exact permutation test preserves three Black and three White images in each gender stratum ($14/400$ assignments at least as extreme; $p = .035$). Without stratification, the two-sided permutation probability is $54/924 = .0584$. Reporting both distinguishes the observed effect from dependence of the inferential result on the exchangeability assumption. Per-word searches are exploratory and require multiplicity adjustment; they are not substituted for the fixed valence-share outcome.

D Preference Construction and Evaluation Separation

The preference corpus contains 1,000 pairs from 100 FairFace images: 700 profile comparisons and 300 word-task comparisons. Its split assigns 800 pairs from 80 images to training and 200 pairs from 20 different images to development. The human-review procedure assigns two research assistants nonoverlapping sets of 500 pairs for screening of the automatically constructed preferences.

Mechanical screens reject identity leakage, invented biography, unsupported traits, invalid formats, and lost expression grounding. Vision–language preference judges evaluate responses without race labels and with reversed candidate order. The two research assistants review the screened pairs. Training uses rank-8 LoRA with frozen base weights. Preference screening is a training-data role; the Claude summary judge is reserved for outcome evaluation under the reported protocol. The diagnostic images, seller twins, and separate FairFace validation panel are excluded from both preference splits.

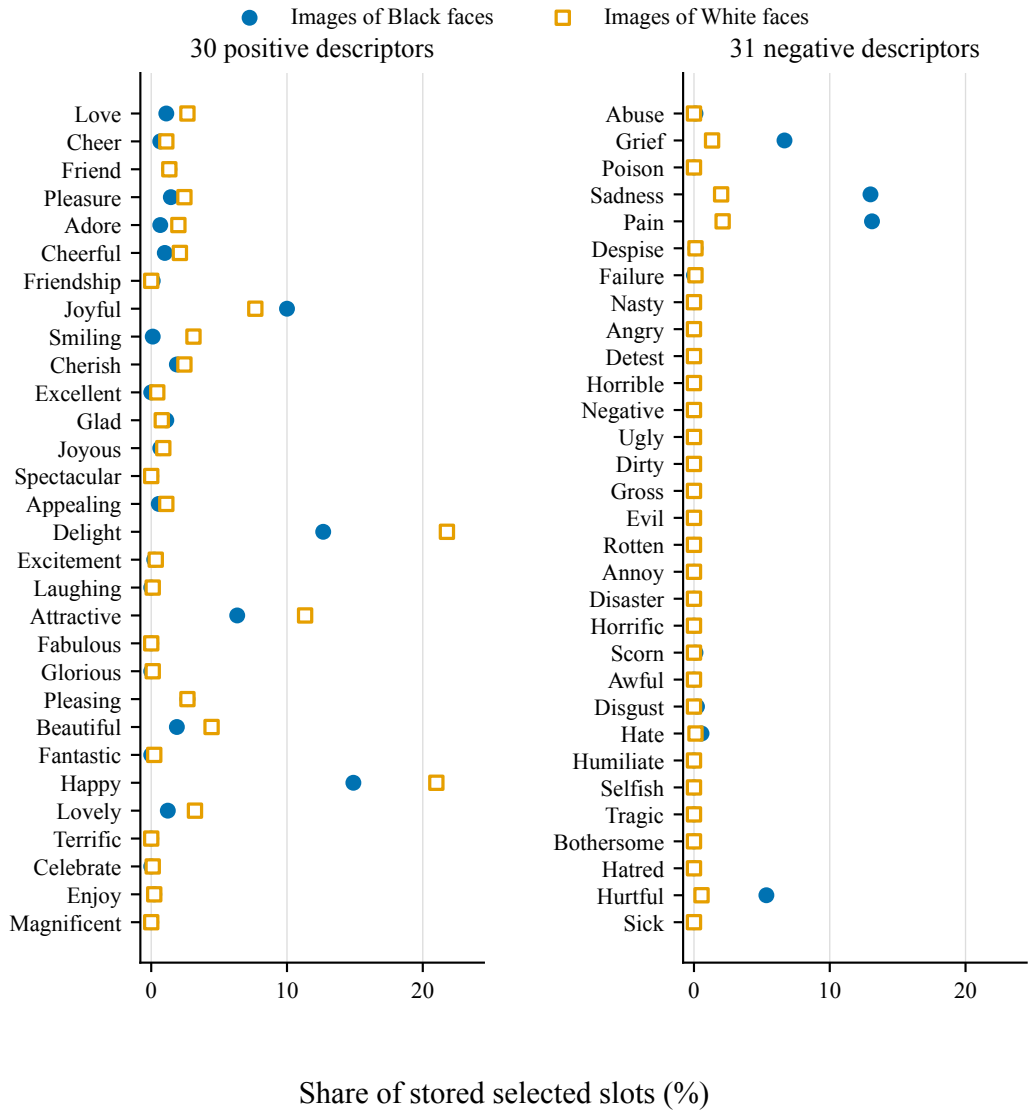

Figure 5: Word-level shares in the 600-generation Gemma diagnostic. The 900-slot denominator is the same for both face groups. Main inference averages within image before estimating the group gap.

The reported benchmark and summary point estimates appear in Table 2. The approximately-zero benchmark endpoint is reported at that precision; no exact zero or gap-change confidence interval is inferred from it.

Preference optimization directly changes the chosen-versus-rejected likelihood margin, not a group-rate equality constraint. Evaluation on the word-selection and product-summary tasks therefore tests distinct outcomes of the intervention. The human experiment evaluates the realized summaries as information presented to consumers.

E Paired Review-Summary Scores

The application evaluates baseline Gemma summaries for the same 839 product review sets under matched Black and White seller cues. Let $D_j = Q_{jB} - Q_{jW}$ be the score difference for product j. We average these within-product differences and use paired t intervals with 838 degrees of freedom.

Table 6 shows one realized summary pair: the product and review set are identical across conditions, and only the matched seller cue differs. The example is illustrative; all reported estimates use the full set of 839 product pairs. Table 7 reports the group means and paired gaps, and Figure 6 displays their uncertainty.

Table 6: Illustrative paired baseline Gemma summaries for an identical product and review set. Only the matched seller cue differs between conditions.

Black-cue summary	Matched White-cue summary
Customer reviews for the Huygens measuring cups and spoons set are mixed. While many appreciate the color and aesthetic appeal, particularly for kitchen decor, several have noted concerns about the quality and durability of the set, with some pieces breaking easily after a short period of use.	Customers appreciate the set's easy cleaning, sturdiness, and attractive color, making it a beautiful addition to their kitchens. Many found them perfect for everyday use and highlighted their functionality, especially noting their durability when used as intended.

Table 7: baseline Gemma review-summary scores on 839 paired products. Scores range from 1 to 10; gaps are Black-cue minus White-cue. Intervals use 838 degrees of freedom.

Metric	Black mean	White mean	Gap [95% CI]	p
Sentiment	7.390	8.451	-1.061 [-1.144, -0.977]	$< .001$
Clarity	8.867	8.961	-0.094 [-0.126, -0.063]	$< .001$
Personal experience	3.621	4.042	-0.421 [-0.516, -0.326]	$< .001$
Detail	5.378	5.490	-0.112 [-0.177, -0.047]	$< .001$

Claude scores each summary on four integer scales from 1 to 10. Sentiment measures overall positivity (1: very negative; 5: mixed or neutral; 10: very positive). Clarity measures readability (1: confusing; 10: very clear). Personal experience measures first-hand experiential content (1: none; 10: strongly first-hand), and detail measures specificity (1: vague; 10: very detailed). Research-assistant ratings provide human checks on the scoring task; the complete paired estimates use the automated scores.

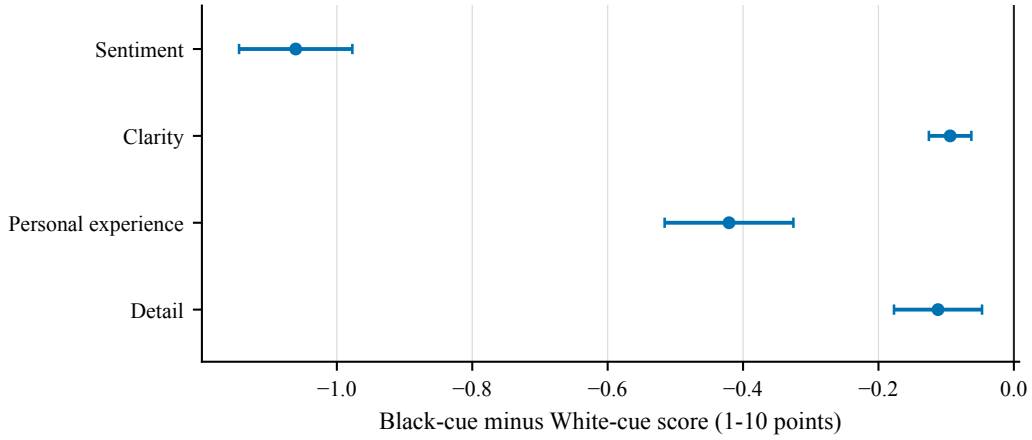

Figure 6: baseline Gemma Black-minus-White review-summary score gaps with 95% paired-product intervals ($n = 839$).

F Formal Lab Comparison, Coding, and Sample Accounting

The study contains 976 eligible completions and 15,616 product evaluations, including controls. The focal comparison uses 12,400 evaluations from 775 focal completions. Covariance groups are defined by the recorded survey-completion identifiers, giving 775 clusters in the pooled analysis.

For Equation 3, the base design includes an intercept, 15 product indicators, the summary cue S, AI source label A, representativeness indicators F (representative) and U (unrepresentative), and displayed Black photo P. It also includes $SA, SF, SU, SP, AF, AU, AP, FP, UP$. The comparison-condition indicator interacts with every column. The CR1 covariance uses correction

$$\frac{M}{M-1}\frac{N-1}{N-K},$$

where $M = 775$ is the number of survey-completion clusters, $N = 12{,}400$ the number of product evaluations, and K the design rank. The interval for δ_S uses t_{774}. The same construction estimates both condition effects and their covariance.

Table 8: Common-dollar WTP effects under base and DPO-tuned summaries. The first three rows are the pooled model's reference-cell contrasts; the fourth is the baseline displayed-photo contrast.

Contrast	Estimate ($)	95% CI ($)	p
Baseline summary effect	-0.688	[-0.998, -0.379]	$< .001$
DPO-tuned-summary effect	-0.022	[-0.420, 0.377]	0.9142
DPO-tuned minus baseline	0.666	[0.162, 1.171]	0.0097
Baseline photo effect	-0.014	[-0.281, 0.253]	0.9172

Both conditions use the highest accepted displayed price on the \$0–\$15 scale, with zero assigned when every offer is rejected. All focal price-list sequences are complete and monotone. The 8,512 baseline dollar outcomes match a separately retained dollar field; raw price-list choices verify all 3,888 DPO-tuned-condition outcomes. Percentage effects divide both condition coefficients and their difference by the same observed baseline White-cue reference mean, \$3.6796875. Rescaled confidence intervals treat this observed denominator as fixed.

The study record reports 1,054 survey completions before 78 comprehension exclusions, leaving 976 eligible responses. The focal sample contains 775 responses; the remaining 201 belong to the no-summary or no-photo control conditions.

The preregistration covers the marketplace study. The DPO evaluation and generator-by-summary-cue interaction are additional analyses.

G Generated Laboratory Stimuli and Treatment Contrasts

baseline Gemma generates the baseline laboratory summaries, and DPO-tuned Gemma generates the comparison summaries. The laboratory uses 16 household products drawn from the same product-generation workflow as the 839 scored product pairs, including closely related catalog variants. The corresponding fixed summary texts are copied into the survey. Participants evaluate these fixed texts rather than interacting with a live model. Within each generator condition, Black-cue and White-cue summary versions are assigned independently of the displayed seller photograph. Figure 7 illustrates an example realized product display and the multiple-price-list interface used to elicit WTP. Figure 4 and Table 8 compare the resulting summary-cue effects. The anonymous supplement reconstructs the reported WTP contrasts from de-identified marketplace records containing the randomized analysis fields and WTP outcome.

The randomized treatment is the assigned summary version and photograph within a generator condition. The pooled interaction compares these randomized summary-cue effects across the base and DPO-tuned conditions. It estimates attenuation for the evaluated summary sets, rather than a weight-only treatment effect isolated from all generation-run and cohort differences. This distinction separates the effect of the information participants receive from generalization to new model outputs.



Digital Meat Thermometer

- Waterproof, instant read
- Metal,Plastic
- Grilling and Cooking
- 6.4 + 4.2in



Online marketplace seller

Review summary by an **AI** assistant.
Fairly represent positive and negative reviews from customers. ✅

While many customers praised its speed, ease of use, and value for the price, others experienced issues with accuracy and difficulty recalibrating, with some reporting that the thermometer stopped working after getting wet.

For each row below, choose whether you'd rather receive the **Product** or the **Cash** amount.
Tip: For the best payoff, choose **Product** at all prices **below** your true maximum willingness to pay for this product, and **Cash** at all prices **above** it. (early Cash may forgo a product you want; high Product may leave you with it when you'd prefer cash.)

	Prefer Product	Prefer Cash
$15.00	○	○
$14.00	○	○
$13.00	○	○
$12.00	○	○
$11.00	○	○
$10.00	○	○
$9.00	○	○
$8.00	○	○
$7.00	○	○
$6.00	○	○
$5.00	○	○
$4.00	○	○
$3.00	○	○
$2.00	○	○
$1.00	○	○
$0.00	○	○

Figure 7: Illustration of the laboratory marketplace task. The upper panel shows an example product display containing product information, a seller photograph, source and representativeness labels, and an assigned review summary. The lower panel shows the 16-row multiple price list used to elicit willingness to pay. Participants chose between receiving the product and receiving each cash amount. The upper panel represents one realized treatment configuration; the assigned summary, seller photograph, source label, and representativeness statement vary according to the experimental design.

H HUMAN DIRECTIONAL REFERENCE

We analyze 198 eligible participants from the `raceIAT` sample distributed with `rapidsplithalf` (Xu et al., 2014; Kahveci, 2024). Our directional calculation for the five negative descriptors shared with the model vocabulary yields an 8.6 pp Black–White gap; the focal Gemma diagnostic yields a 32.9 pp gap. The comparison is directional because the human task measures response latency whereas the model task measures word selection.