Investigating Demographic Distribution Shifts in Big Five Personality Scoring by Claude Sonnet 4.6

Piper Fleming
Department of Computer Science
Stanford University
Stanford, CA 94305
piperf@stanford.edu

Carolyn Smith
Department of Computer Science
Stanford University
Stanford, CA 94305
carsmith@stanford.edu

Abstract

Large language models implicitly infer demographic and psychological characteristics when reasoning about text authored by individuals, yet it remains unclear whether explicitly providing demographic information systematically biases the personality profiles they attribute to those individuals. We investigate this question using the PANDORA dataset of Reddit users with self-disclosed age, gender, and personality information. We conduct a within-subject experiment in which identical comment sets are evaluated by Claude Sonnet 4.6 under four prompt conditions that vary only in the demographic information provided to the model. Analyzing 497 users, we find that (1) making demographics explicit shifts predictions even when the same information is inferable from text; (2) gender labeling produces a uniform shift across all five traits, with direction reversing cleanly by swap direction; and (3) age labeling produces trait-specific shifts consistent with the maturity principle in personality psychology. A calibration diagnostic reveals that these effects operate as near-additive offsets that preserve user rank order rather than as trait-specific stereotype-driven reorderings. These findings suggest that explicit demographic cues can bias personality assessments independently of the textual evidence on which those assessments are based, raising important considerations for fairness and evaluation in automated personality profiling systems.

1 Introduction

Human personality can be described along five orthogonal dimensions known as the Big Five [1]: *Openness* (imagination, intellectual curiosity, receptiveness to new ideas and experiences), *Conscientiousness* (organization, dependability, self-discipline, goal-orientation), *Extraversion* (sociability, talkativeness, assertiveness, positive emotionality), *Agreeableness* (trust, altruism, cooperation, empathy), and *Neuroticism* (tendency toward negative emotions, emotional instability, anxiety), all scored from 0 to 100. These traits have been shown to be particularly associated with demographic stereotypes like age and gender [3]. For instance, women are perceived as higher on agreeableness and conscientiousness [3], while conscientiousness has been found to increase and neuroticism to decrease with age [4].

Large language models are increasingly being used for tasks that require inferring psychological attributes from text, including hiring screens, behavioral analytics, and mental-health triage. If LLMs encode demographic stereotypes in their personality predictions, their outputs can systematically disadvantage individuals based on group membership rather than actual behavioral evidence. Understanding the structure of such biases, whether they are trait-specific, uniform, or calibration-based, is essential for auditing and mitigating harm in these high-stakes applications.

CS 281, Spring 2026

In this work, we examine how demographic stereotypes emerge in LLM-based personality evaluations inferred from Reddit user comments. We use the PANDORA dataset [2], a large-scale collection of 17.6 million comments from 10.3k Reddit users with self-disclosed demographic information. By varying only a single-line demographic header in an otherwise identical prompt, we isolate the causal effect of specifying a user's age and gender on Claude Sonnet 4.6's Big Five scores. Our within-subject design ensures that any observed shifts are attributable to the demographic labeling rather than to differences in the comment text.

2 Methods

The full comment corpus was obtained from the PANDORA authors [2] via their distribution request form. No verbatim comments, usernames, or identifying information appear in this paper. Out of the 10.3k users, a stratified sample was constructed using the following criteria.

Limitation: gender is treated as binary (m/f) throughout; the 143 users reporting gender "t" were excluded to enable a clean swap, which limits generalizability to non-binary individuals.

- Eligibility: users who were clearly reported as male (m) or female (f), and who had an age that qualified as "young" (less than 22) or "older" (greater than 30). This distinction, as well as the gap, was chosen to ensure that we were comparing actually contrastive datapoints.
- Stratification: 500 random users were then chosen, with 125 coming from each gender x age group. This ensured that we were measuring contrast within a balanced subgroup, since computing with the full eligible pool would have an over-representation of young males.
- Comment content and quantity: several users had a lack of comments that were either too short (defined as less than 5 words), or had content that would lead to signal leakage (more on this below).

For each user in the chosen subgroup, the following filters were applied:

- Only comments in English were considered.
- No signal leakage. For our purposes, signal leakage is defined as a comment that contains some clear information about the user's demographics that skews the response of the LLM. As an example, we saw both lists of relevant subreddits and the presence of a word relating to personality in the comment. To detect the presence of the latter, a regex scrub of common terms was applied to prevent the LLM from directly reading these characteristics from the text.

Of the almost 900k author-matched comments, 714k were left to sample from. The final corpus that was fed to the LLM had 14.5k comments across 497 users, with a median of 30 per user (23 users had less than 30 after filtering was applied).

Within each user, we applied a four condition design, where comments are identical across conditions for each user and only a single-line demographic header in the prompt is varied.

The four conditions are as follows:

- C0: no demographic header (control)
- C1: real demographics (User profile: {age}-year old {gender})
- C2g: real age, gender label flipped
- C2a: real gender, age label swapped (ages under 22 are set to 45; ages over 30 are set to 20).

And three shifts are specifically measured per user:

- C0 - C1: effect of making demographic information explicit vs. leaving the model to infer it from text.
- C1 - C2g: pure effect of the gender label (text and age held constant).
- C1 - C2a: pure effect of the age label (text and gender held constant).

This methodology was chosen because any systematic shift between C1 and C2x must therefore come from the label. Figure 1 summarizes the full data and analysis pipeline.

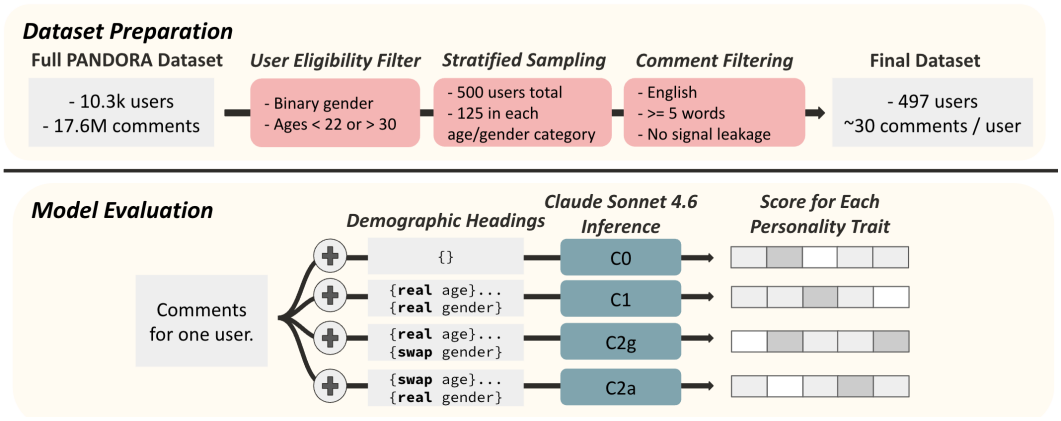

Figure 1: Overview of the experimental pipeline, from PANDORA dataset sampling through prompt condition design and LLM inference.

LLM Inference For our model integration, we used Claude Sonnet 4.6 [5], with a temperature of 0 and random seed of 281. The model was accessed via the Anthropic API using a single-turn prompt per call. The system prompt read: *"You are an expert psychologist assessing personality from social media writing. Below are comments written by a single Reddit user."* An optional demographic header line (`User profile: {age}-year-old {gender}.`) was inserted before the comments in conditions C1, C2g, and C2a. The model was then asked to rate the user on each Big Five trait on a 0–100 integer scale and return a single JSON object with scores and a one-sentence justification per trait.

Statistical Analysis The unit of analysis is the per-user score shift, $\Delta_i = s_i^{\text{treatment}} - s_i^{\text{baseline}}$, with each user serving as their own control. For each (contrast, trait) pair, we apply a Wilcoxon signed-rank test, a nonparametric paired test that does not assume normality. To account for multiple comparisons, we apply a Bonferroni correction across $m_{\text{traits}} \times m_{\text{contrasts}} = 5 \times 3 = 15$ tests, yielding a corrected significance threshold of $\alpha^* = 0.05/15 \approx 0.0033$. Effect size is reported using paired Cohen's $d = \bar{\Delta}/\text{SD}(\Delta)$, where $\bar{\Delta}$ is the mean score shift and $\text{SD}(\Delta)$ its standard deviation across users. We additionally report 95% bootstrap confidence intervals based on 2,000 resamples. Because gender and age swaps are anti-symmetric by construction ($\text{m} \to \text{f}$ is expected to somewhat mirror $\text{f} \to \text{m}$), effects may cancel when pooled across directions. We therefore perform a directional analysis stratified by swap direction ($\text{m} \to \text{f}$ vs. $\text{f} \to \text{m}$; young→older vs. older→young).

Calibration-vs-Stereotyping Diagnostic To determine whether label-induced shifts arise from (a) trait-specific stereotype-driven reordering of users or (b) uniform additive recalibration, we compute three quantities per (trait, swap, direction): (i) Spearman ρ between baseline and swapped scores; high ρ indicates preserved rank order (calibration); low ρ indicates reshuffling (stereotyping); (ii) ratio of per-user shift SD to baseline SD where a small ratio indicates a near-additive shift; (iii) Ordinary Least Squares regression slope of swapped-on-baseline; a slope ≈ 1 with nonzero intercept is the signature of a pure additive offset.

3 Experiments & Sociotechnical Analysis

Experimental Methodology Overview We run four experiments using the same within-subject infrastructure described in §Methods: each user contributes scores under four prompt conditions (C0, C1, C2g, C2a), and we analyze per-user shifts between specified pairs of conditions. Experiments 1–3 test whether specific demographic manipulations shift the model's Big Five predictions. Experiment 4 is a diagnostic that decomposes any observed shift into "uniform calibration" vs. "trait-specific stereotyping" components. Significance criterion used were paired Wilcoxon, Bonferroni-corrected α = 0.0033, and the effect size reported as paired Cohen's d.

3.1 Experiment 1: Effect of Making Demographics Explicit

Null Hypothesis Adding an explicit demographic header to the prompt will not significantly shift the model's Big Five score predictions, on the grounds that demographics are already inferable from the comment text. Equivalently, the mean per-user shift, $\bar{\Delta}$, in each of the five personality scores is zero.

Setup Contrast C0 (no demographic header) vs. C1 (real demographics in header), holding user and comments fixed.

Results All five traits show statistically significant shifts after Bonferroni correction (p ranging from 10^{-5} to 10^{-15}). Mean signed shifts: openness $+0.94$, conscientiousness $+0.47$, extraversion -0.91, agreeableness $+0.59$, neuroticism $+1.11$. Cohen's d ranges from 0.19 to 0.40. Appendix Figure 3 shows a heatmap of mean signed shifts across all contrasts and traits.

Technical Interpretation The null hypothesis is rejected: Making demographics explicit shifts model predictions in a statistically significant way.

Socio-Technical Interpretation Adding explicit demographic information systematically changes personality predictions, even when the underlying comment text is unchanged. The same individual, represented by the same comments, receives different personality scores depending on whether demographic information is explicitly supplied. This indicates that model outputs are sensitive to the representation of demographic attributes in the prompt, not solely to the textual evidence itself. For systems that use personality inference in downstream decision-making, such sensitivity may complicate efforts to ensure consistency and fairness across deployments, particularly when demographic information is available in some contexts but omitted in others.

3.2 Experiment 2: Effect of Varying the Gender Label (Headline Experiment)

Null Hypothesis Flipping the gender label in the prompt, while holding comment text constant, produces no systematic shift in Big Five predictions: $\mathbb{E}[\Delta_i^{(t)}] = 0$ for each trait t.

Setup Contrast C1 (real demographics) vs. C2g (flipped gender, age held constant) across all 497 users. We run two analyses: *aggregate* (all users pooled) and *directional* (split by real gender: m→f, $n = 248$; f→m, $n = 249$). The directional analysis is necessary because opposite swap directions produce opposing shifts that partially cancel when pooled, attenuating the true effect.

Results *Aggregate*: Only neuroticism survives Bonferroni correction (mean shift -0.59, $d = -0.20$); remaining traits show $|\bar{\Delta}| \leq 0.31$, suppressed by cancellation across swap directions.

Directional: Both directions show large, significant effects on all five traits (all $p < 10^{-3}$). The m→f swap raises every trait: agreeableness $+1.94$ ($d = 0.83$), conscientiousness $+1.91$ ($d = 0.80$), extraversion $+1.08$, neuroticism $+0.56$, openness $+0.47$. The f→m swap lowers every trait with roughly similar magnitudes: conscientiousness -2.43 ($d = -1.00$), neuroticism -1.73, extraversion -1.53, agreeableness -1.31, openness -0.44. The two directions are approximately mirror images, suggesting the model applies a symmetric gender prior. Figure 2 shows per-user shift distributions. Appendix Figure 4 illustrates the symmetry.

Technical Interpretation The null hypothesis is rejected, but the effect structure is unexpected. Löckenhoff et al. [3] find that human gender stereotypes are trait-specific. For example, women perceived higher on agreeableness and conscientiousness but with opposing effects within neuroticism and extraversion. Instead, we observe a uniformly-signed shift, where every Big Five trait moves in the same direction with the same gender swap. This is inconsistent with a trait-by-trait stereotype model.

Socio-Technical Interpretation The model exhibits a clear gendered bias, but not one that is unidirectional in its outcomes. Female labeling raises traits read as positive in workplace contexts (conscientiousness, agreeableness) *and* neuroticism, which is read as negative. Downstream harm therefore depends on which traits the deployment system rewards. For example, a hiring screen weighting conscientiousness positively would favor users labeled female, while one penalizing neuroticism could disadvantage them. The bias is bidirectional and structurally consistent.

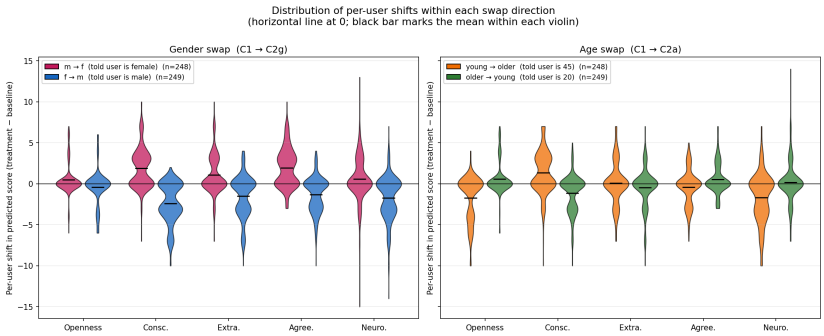

Figure 2: Distribution of per-user shifts within each swap direction. The horizontal line marks zero. The black bar marks the mean within each distribution. In the left plot, swapping user genders leads to greater distributional separation for some traits (conscientiousness, agreeableness) than for others (openness). Age swap distributions show trait-specific asymmetry.

3.3 Experiment 3: Effect of Varying the Age Label

Null Hypothesis Flipping a users age bin (older/younger) in the prompt, while holding comment text constant, produces no systematic shift in Big Five predictions: $\mathbb{E}[\Delta_i^{(t)}] = 0$ for each trait t. We expect (informed by personality-psychology literature on the "maturity principle") that older labeling will be associated with higher conscientiousness, lower neuroticism, and possibly lower openness.

Setup Contrast C1 (real demographic header) vs. C2a (gender same, age swapped). If a user's real age is under 20, they swap to 45, if their real age is over 30, they swap to 20. We again conduct aggregate and directional analyses.

Results *Aggregate:* Openness (-0.58, d = -0.22, p $\approx 10^{-6}$) and neuroticism (-0.77, d = -0.26, p $\approx 10^{-8}$) show significant shifts after Bonferroni. These aggregate effects are non-zero, indicating that the two swap directions do not cancel symmetrically.

Directional (young→older, n=248): Lower openness (-1.71, d = -0.63), higher conscientiousness (+1.35, d = 0.52), lower neuroticism (-1.70, d = -0.56).

Directional (older→young, n=249): Mirror pattern, smaller for some traits: higher openness (+0.56), lower conscientiousness (-1.16).

From Figure 4 (right panel), age effects are trait-specific, with different traits moving in different directions, unlike the gender swap, where every trait moved uniformly.

Technical Interpretation The null hypothesis is rejected with trait-specific structure. Age-label swaps seem to produce selective effects consistent with the maturity principle: older labeling is associated with lower neuroticism and lower openness, matching expected developmental trajectories. That age and gender produce qualitatively different shift structures suggests the model encodes these two demographic signals via distinct mechanisms.

Socio-Technical Interpretation Age-related shifts in personality scores plausibly reflect empirical patterns corroborated in the psychology literature, which raises a normative question: is a model that adjusts personality predictions based on age applying valid prior information, or stereotyping individuals based on group statistics? In clinical or conversational contexts, age-based adjustment may be defensible, while in employment decision-making, it likely constitutes age discrimination regardless of empirical validity.

3.4 Experiment 4: Calibration vs. Stereotyping Diagnostic

Hypothesis The label-induced shifts observed in Experiments 2 and 3 could arise via (a) trait-specific stereotyping (the model reshuffles which users are predicted high vs. low on each trait based on the label) or (b) uniform calibration shift (the model preserves user rank order but slides all predictions up or down by a roughly constant amount). We test which mechanism is dominant.

Setup For each (trait × swap × swap-direction), we compute (i) Spearman rank correlation between baseline and swapped scores, (ii) the ratio of per-user shift SD to baseline SD, and (iii) the slope of

an OLS regression of swapped on baseline scores. Pure stereotyping predicts low ρ, pure calibration predicts $\rho \approx 1$, slope ≈ 1, and small SD ratio.

Results All three diagnostics consistently support a calibration-shift interpretation across both swaps and all five traits. Spearman ρ ranged 0.93–0.99, SD ratios 0.14–0.33, regression slopes 0.93–1.07 with mean shifts captured almost entirely in the intercept. Appendix Figure 5 shows that within each (trait, swap-direction) cell the per-user predictions in the swapped condition fall almost perfectly on a line parallel to $y = x$ but offset.

Technical Interpretation The gender and age effects from Experiments 2 and 3 operate as near-uniform calibration shifts rather than as label-conditioned reordering of users. The model may not be encoding the stereotype that "women are higher on agreeableness". Rather, it could be using the gender label to adjust the reference scale on which all users are rated.

Socio-Technical Interpretation This distinction has consequences for bias mitigation. Stereotyping bias might be addressed by training interventions that decouple specific traits from demographic labels. Calibration-shift bias is harder to address through such targeted interventions because the model's entire reference distribution differs by group. Mitigation likely requires either post-hoc rescaling of outputs per demographic group (raising its own fairness concerns) or fundamental changes to how the model represents demographic information internally.

General Observation for All Demographic Swaps Qualitative inspection of the model's per-call justifications reveals a consistent mechanism: the model attributes the *same observed behaviors* to different traits depending on the demographic label. In the gender swap, hostility in a user's Reddit comments is described as "elevated neuroticism" under the male label is reattributed to "low warmth" under the female label. In the age swap, goal-tracking described as "strong self-discipline" under the young label becomes "stress-driven control" under the older label.

4 Conclusion

Using a within-subject design on 497 PANDORA users, we found that explicit demographic headers shift Big Five personality score predictions in a statistically significant way. Further, we demonstrate that gender labeling produces a uniform signed shift across all five traits (reversing cleanly with the swap direction), inconsistent with the trait-specific structure documented in human gender stereotype literature [**?**]. Age labeling produces trait-specific shifts partially consistent with the maturity principle [4]. A calibration diagnostic showed these effects are near-additive offsets in score distributions for each trait, that preserve user rank order rather than produce trait-specific reorderings.

Results should be interpreted with caution. The PANDORA dataset relies on Reddit users who opt-in to share their data. This self-selecting group is unlikely to be fully representative of personality variance found in the general population. Reddit users skew young, male, and Western, which may interact with the demographic swap effects observed here in ways that are difficult to disentangle.

Future work should replicate across other LLMs, demographic axes (ex. race, nationality), and text domains, and examine whether post-hoc demographic-stratified rescaling can reduce calibration shift without introducing new fairness concerns.

Acknowledgements

Analysis code for this project was written with the assistance of Claude (Anthropic), used as an AI coding tool throughout the project.

References

[1] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of Personality*, vol. 60, no. 2, pp. 175–215, Jun. 1992. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

[2] M. Gjurković, V. M. Karan, I. Vukojević, M. Bošnjak, and J. Šnajder, "PANDORA Talks: Personality and Demographics on Reddit," in *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, Online, Jun. 2021, pp. 138–152. https://doi.org/10.18653/v1/2021.socialnlp-1.12

[3] C. E. Löckenhoff et al. Gender stereotypes of personality: Universal and accurate? *Journal of Cross-Cultural Psychology*, 45(5):675–694, 2014. doi: 10.1177/0022022113520075.

[4] B. W. Roberts and D. Mroczek, "Personality trait change in adulthood," *Current Directions in Psychological Science*, vol. 17, no. 1, pp. 31–35, Feb. 2008. `https://doi.org/10.1111/j.1467-8721.2008.00543.x`

[5] Anthropic, "Claude Sonnet 4.6 Model Card," Anthropic, 2024. `https://www.anthropic.com/claude`

A Supplementary Figures

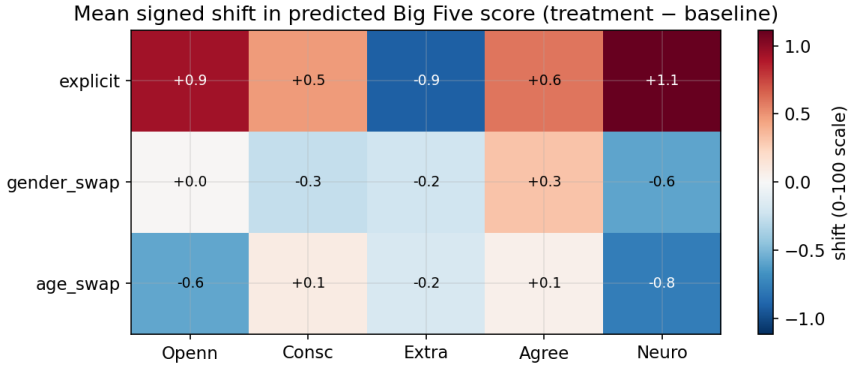

Figure 3: Mean signed shift in predicted Big Five score (treatment − baseline) across all three contrasts and five traits. Red indicates positive shift, blue indicates negative shift. The *explicit* row shows consistent non-zero shifts; the *gender_swap* row masks directional effects that cancel in aggregate; the *age_swap* row reveals trait-specific patterns.

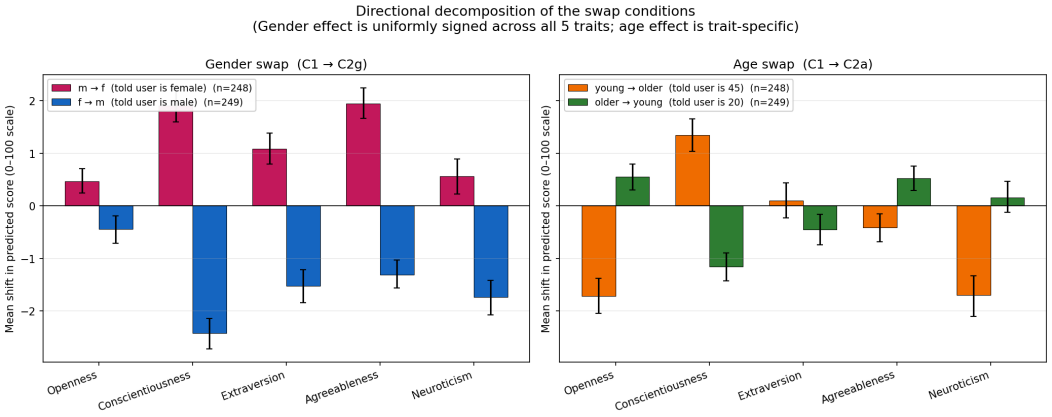

Figure 4: Directional decomposition of the gender swap (left) and age swap (right). Each bar shows the mean per-user shift with 95% bootstrap confidence intervals. Gender labeling produces a uniformly signed shift across all five traits; age labeling produces trait-specific shifts largely consistent with the maturity principle.

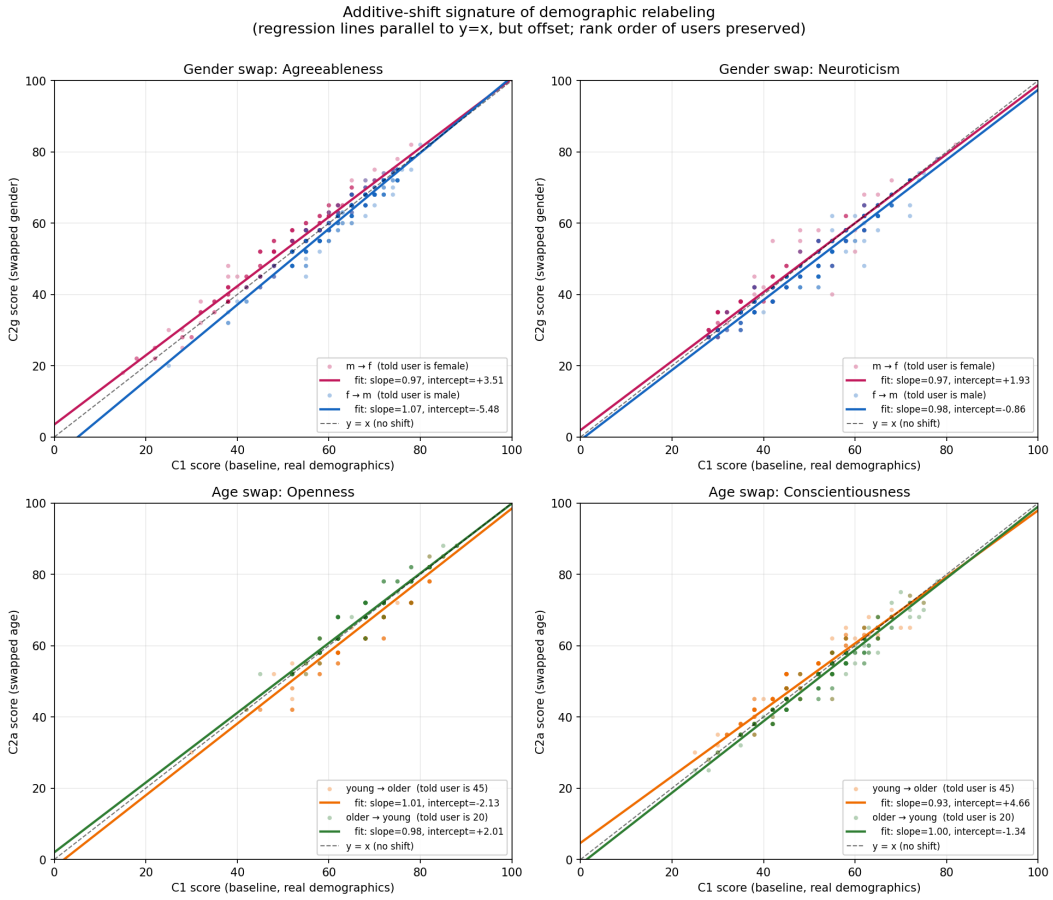

Figure 5: Additive-shift signature of demographic relabeling. Each panel plots baseline score (x-axis) against swapped score (y-axis) for a representative (trait, swap) combination. Regression lines run nearly parallel to $y = x$ but with a nonzero intercept, indicating that rank order is preserved and the demographic label acts as a near-uniform additive offset.