Can a Fine-Tuned Language Model Assess Personality Equitably Across Race in Older Adults?

Tu Do, Tong Li, Patrick L. Hill, Seanna C. Leath, Mehak Gupta, and Joshua R. Oltmanns

Author's note:

Tu Do, M.A., Department of Psychological & Brain Sciences, Washington University in St. Louis; Tong Li, M.S., Department of Computer Science and Engineering, Washington University in St. Louis; Patrick L. Hill, Ph.D., Department of Psychological & Brain Sciences, Washington University in St. Louis; Seanna C. Leath, Ph.D., Department of Psychological & Brain Sciences, Washington University in St. Louis; Mehak Gupta, Ph.D., Department of Computer Science, SMU; Joshua R. Oltmanns, Ph.D., Department of Psychological & Brain Sciences, Washington University in St. Louis.

This research was supported by the NIH under Award Numbers R01-AG061162, R01-AG045231, and R01-MH077840. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

The authors would like to thank researchers and participants from the St. Louis Personality and Aging Network (SPAN) for their dedicated efforts in collecting the data for this project.

Black/White Differences in Language-Based AI Modeling of Personality

Correspondence should be addressed to Tu Do, Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO. Email: do.t@wustl.edu

Black/White Differences in Language-Based AI Modeling of Personality

Abstract

Artificial intelligence (AI) enables large-scale personality assessment from language. Differential performance across racial groups should be evaluated before applied use. The present study systematically examines bias in language-based personality prediction from life narrative interviews across Black and White American community older adults. Linguistic Inquiry and Word Count (LIWC), BERTopic, and fine-tuned RoBERTa language model predictions of FFM traits were evaluated in a sample of n = 450 Black and n = 450 White older adults matched on education and income. LIWC and BERTopic associations differed little across race, with the few significant differences near chance levels. Fine-tuned RoBERTa predictive accuracy differed significantly for extraversion: the model performed well for White and near chance for Black participants (r = .41 vs. .17). Calibration likewise differed significantly across race only for extraversion. Openness showed significant directional bias, underpredicting scores for Black participants relative to White participants, without a significant calibration difference. These findings indicate that strong aggregate performance across race can mask differential subgroup performance.

Keywords: language, natural language processing (NLP), artificial intelligence (AI), personality, five-factor model, big five, racial differences

Can a Fine-Tuned Language Model Assess Personality Equitably Across Race in Older Adults?

The Five-Factor Model (FFM) of personality originated from the lexical hypothesis, which states that personality traits are encoded in language (Allport & Odbert, 1936; Galton, 1884). Over the past one hundred years, a large body of literature has supported the validity of five broad domains of personality (neuroticism, extraversion, openness, agreeableness, and conscientiousness) that emerged from research by Allport on the English dictionary (Goldberg, 1993; John, 2021). More recently, advances in artificial intelligence (AI) make it possible to return to a lexical focus in personality assessment (J. R. Oltmanns et al., 2026; Park et al., 2015). However, racial fairness of language-based AI personality assessments remains largely untested.

The majority of personality assessment research relies on self-report, which has documented limitations (Paulhus & Vazire, 2007). For example, people may respond in socially desirable ways untrue to their real personality traits or have limited insight into their own personality (Vazire, 2010). Supporting these limitations are meta-analyses demonstrating that self-other agreement on personality traits varies across traits and raters (Connelly & Ones, 2010). As a result, psychologists strive to incorporate multimethod assessment in research and clinical practice to improve validity (APA Task Force on Psychological Assessment and Evaluation Guidelines, 2020). AI-based assessment methods may provide a much longed-for scalable and more behavioral method of integrating multimethod assessment into research and practice (Brickman et al., 2025; Kjell et al., 2024).

A growing literature has established that personality can be predicted from natural language and other digital traces. Early work showed that Big Five traits could be inferred from social media language and digital footprints at correlations in the .30 to .40 range (Park et al., 2015), effects that a subsequent meta-analysis found to be robust across platforms and trait

domains (Azucar et al., 2018). More recent work has extended language-based assessment to video interviews (Hickman et al., 2021), smartphone-sensed behavior (Stachl et al., 2020), and life narrative interviews (Hussain et al., 2026; Oltmanns et al., 2026), and large language models can now infer traits from text without task-specific training (Peters & Matz, 2024). The appeal of these methods is practical: personality can be scored from behavior people generate anyway, without additional participant burden, at a scale and cost that self-report cannot match, and in settings—clinical intake, longitudinal cohorts, archival text—where administering a 240-item inventory is not feasible.

Just as self-report measures require validation across diverse populations, so do AI-based assessments. Algorithmic systems have produced documented racial disparities in high-stakes settings, including a widely used health care algorithm that systematically underestimated the care needs of Black patients (Obermeyer et al., 2019). Relatedly, the corpora used to pretrain language models can overrepresent hegemonic viewpoints and the language of already-privileged groups (Bender et al., 2021), and several NLP systems have shown poorer performance on text containing features associated with African American English (Blodgett et al., 2020). AI-based personality assessments should therefore be evaluated for validity, reliability, calibration, and differential performance in samples that adequately represent Black Americans before their scores are used to draw substantive conclusions in research or inform clinical decisions.

Several mechanisms could produce cross-group differences in how personality is expressed in language: sociolinguistic variation in language use, such as the distinct grammatical and rhetorical conventions of African American English (Green, 2002; Rickford, 1999; Smitherman, 1977); cultural variation in narrative themes and forms in life stories (McAdams &

Pals, 2006; Turner et al., 2022); and systematically different life experiences including unequal life-course exposure to discrimination and differences in socioeconomic resources and opportunities (Ferraro & Shippee, 2009; Williams & Mohammed, 2009). These differences do not imply inherent racial differences in personality. Rather, they suggest that the same personality trait may be expressed through different linguistic, autobiographical, or contextual features across groups. If language-based AI models learn associations between personality scores and linguistic features from one group more strongly than another, then apparently accurate models in the full sample may still perform differently across racial groups.

The present study focuses on older adults, a group in whom personality traits are clinically and practically meaningful, relating to physical health, caregiving relationships, and quality of life (Chapman et al., 2011; Mroczek & Spiro, 2007; Terracciano et al., 2008). A scalable, behaviorally grounded assessment method would therefore be especially valuable in this group. Yet older adults are also underrepresented in AI and natural language processing research, including in the text corpora used to pretrain large language models, so generational language patterns, historical references, and narrative conventions characteristic of older adults may be less well represented in pretrained model parameters (Bender et al., 2021). The present study therefore examines the intersection of two underrepresentations: race and age—in a cohort whose formative experiences included segregation and the Civil Rights movement.

Fairness in algorithmic prediction refers to whether a model performs equitably across groups, rather than systematically advantaging or disadvantaging some groups over others (Barocas et al., 2019; Mehrabi et al., 2021). In the context of psychological assessment, an unfair model is one whose predictions are less accurate, or systematically biased, for one group than another. This concern has been formalized psychometrically as machine-learning measurement

bias (MLMB), defined as differential functioning of a trained model across subgroups (Tay et al., 2022). MLMB manifests in two ways: when a model produces different predicted score levels for individuals with the same ground-truth score, and when it yields different predictive accuracy for the scores the model was trained to predict across subgroups. This framing connects algorithmic fairness to the measurement-invariance tradition in psychological assessment and motivates evaluating language-based personality models across racial groups before applied use. Fairness is not a single property, but a family of criteria: For example, a model may show comparable average error across groups yet still differ in how well it rank-orders individuals, how well it is calibrated, or whether it systematically over- or underpredicts (Chouldechova, 2017; Kleinberg et al., 2016). No model can satisfy all fairness criteria when groups differ in the outcome distribution, so fairness is best understood as a profile of indicators. We use *equitable* in this measurement sense: whether the trained model functions equivalently across Black and White older adults, assessed as a profile of complementary criteria. We do not address the broader social question of whether personality-based assessment produces equal outcomes across groups (Tay et al., 2022).

Empirical tests of demographic fairness in language-based mental-health prediction remain rare. Aguirre et al. (2021) trained depression classifiers on two Twitter datasets in which depression was derived from users' public self-disclosures of a diagnosis, while race/ethnicity and gender were inferred using demographic classification models. Classifiers performed worse for people of color overall, although which subgroup fared worst varied across the two datasets. The disparity persisted even when each demographic group contributed equal numbers of users to the training data.

Rai et al. (2024) offer a more direct test of the underlying question, using self-reported race and self-reported depression severity. They systematically tested fairness in language-based AI prediction of depression across Black and White American adults. In this case, 868 adults (n = 434 White, n = 434 Black) consented to share Facebook status updates, from which depression was modeled. In terms of linguistic features associated with depression, several topics (e.g., worthlessness) did not equally relate to depression for Black participants compared to White participants. Most notably, even the most widely cited linguistic feature of depression ("I" pronoun usage) did not transfer to Black participants. Machine learning models were trained separately within the Black and White subsamples, with training data held constant, and crossed in a 2 x 2 train-by-test design using two feature sets (n-grams plus LIWC categories and Latent Dirichlet Analysis topics; BERT embeddings). Prediction was consistently stronger in White test sets (r = .16 to .39) than in Black test sets (r = .06 to .13). Training exclusively on Black participants' language did not improve prediction for Black participants (r = .13 and .06 across the two feature sets); the models performed poorly for Black participants regardless of which group they were trained on. This indicates that language features identified at the aggregated level—even some of the strongest—may not relate to psychological constructs equally across racial groups.

The present study examines racial fairness in personality prediction from life narrative interviews in Black (n = 450) and White (n = 450) older adults in the St. Louis Personality and Aging Network (SPAN) study (total N = 900). These data come from the same study as Oltmanns et al. (2026), who examined personality prediction from language aggregating across race. The RoBERTa language model (Liu et al., 2019) was fine-tuned (i.e., retrained) on the language from the life narrative interviews to predict FFM personality scores derived from the

NEO-Personality Inventory-Revised (Costa & McCrae, 1992). The fine-tuned RoBERTa language model predicted personality in the test data at relatively large effect sizes: R^2 ranged from .10 (agreeableness and conscientiousness) to .20 (openness), with an average of .15 (extraversion). Pearson *r* correlation between the fine-tuned RoBERTa personality predictions and actual personality scores ranged from $r = .33$ (agreeableness and conscientiousness) to $r = .43$ (neuroticism), with an averageof .41 (extraversion and openness). These effect sizes are remarkable because they approach a large effect size for convergent validity correlations between two self-report scale scores of the same psychological construct. However, despite the sample containing a large proportion of Black participants, systematic testing of racial differences was outside the scope of the paper's research question.

To our knowledge, no prior study has directly examined racial fairness in language-based AI models of personality. We address three questions: First, do linguistic features and narrative topics associated with personality differ between Blak and White older adults? We examine this question exploratorily using LIWC categories and BERTopic-derived topics. Second, does the predictive performance of the fine-tuned RoBERTa model differ across Black and White participants, and does performance depend on which racial group the model was trained and tested on? Third, beyond difference in predictive performance, does the model function equitably across Black and White participants—that is, does it show comparable error magnitude, directional bias, and calibration? Together, these analyses examine racial fairness in language-based AI personality assessment across complementary levels.

Method

The present study was exploratory and was not preregistered. The results should be replicated before strong conclusions are drawn. Code for the analyses is available online

(https://anonymous.4open.science/r/RacialDifferences-01E1/). Personality data are available on the Open Science Framework (https://osf.io/6pq7w/view_only=651a190683d94e8e90ed4cd78ef7ce2f). However, life narrative interviews are not freely available due to potential privacy concerns.

Procedure

The St. Louis Personality and Aging Network (SPAN) is a longitudinal study of a representative community sample of 1,630 older adults recruited across the St. Louis area. Target research participants were identified for participation using contact of listed phone numbers and the Kish (1949) method within households. Once in the laboratory, participants completed life narrative interviews, the NEO-Personality Inventory-Revised (NEO-PI-R), and other self and informant-report measures of personality and health (T. F. Oltmanns et al., 2014). The protocol was approved by the university institutional review board.

Participants

Participant demographic information is presented in Table 1. N = 1,409 participants completed the life narrative interview along with personality measures. Recruitment and demographics are described in detail elsewhere (T. F. Oltmanns et al., 2014; Spence & Oltmanns, 2011). Race and ethnicity were representative of the St. Louis area. Participants came from a wide range of socioeconomic statuses, with a somewhat higher median household income compared to the median in St. Louis at the time the data were collected. After initially lower participation rates compared to Black women and White men and women, Black men were successfully oversampled (Spence & Oltmanns, 2011). Initial sample consisted of $N_{\underline{\text{white}}}$ = 913 ($M_{\text{age\_White}}$ = 59.8, SD = 2.8) and N_{Black} = 450 ($M_{\text{age\_Black}}$ = 59.7, SD = 2.7). Gender proportions

are comparable across the two races, with 54.4% and 56.9% being female in the White and Black sample, respectively.

Matching

Participants whose race was not Black or White were excluded, resulting in 1,363 individuals. We performed propensity score matching using one-hot encoded education level and annual household income to balance key socioeconomic covariates because they shape vocabulary, syntactic complexity, and discourse conventions. Education and income were selected as matching variables deliberately, to ensure that observed group differences are not driven by socioeconomic confounds. We fit a logistic regression model predicting race group membership from one-hot encoded education level and annual household income to estimate each participant's propensity score, and then performed 1:1 nearest-neighbor matching without replacement, with a caliper of 0.45 on the raw propensity-score scale (0–1). This relatively wide caliper was chosen deliberately to retain the full Black sample given limited socioeconomic overlap between groups. This resulted in 450 matched Black-White pairs, with M_{age_White} = 60.1 SD = 2.9 (58.4% female). However, participant sample sizes in both groups vary across traits due to missing values.

Matching substantially shifted the socioeconomic composition of the White sample to resemble the Black sample. In the full White sample (N = 913), over half of participants held a bachelor's degree or higher and the income distribution had a heavy upper tail (17.1% earning $140,000 or more). After matching, the reduced White sample (N = 450) shifted downward to resemble the Black sample, in which "some college" became the modal education category (25.8%, vs. 24.9% in the Black sample) and the high-income tail shrank substantially (1.1% earning $140,000 or more). Age ($M_{age_reduced\ White}$ = 60.5, M_{age_Black} = 59.7) and gender (58.2% vs.

56.9% female) were comparable across the matched groups. The reduced White sample is therefore well matched to the Black sample on the targeted socioeconomic variables but is not representative of the original White sample.

Measures

Life Narrative Interview

We used a reduced version of the life story interview that was developed by McAdams (1993). Each participant provided their life story beginning at age 18, divided into three or four chapters. They were then asked to name the best and worst characters in their life story, high and low points, and a turning point. On average, these interviews lasted 20 minutes. The entire narrative responses of the interviewees were used in analysis. Black participants' narratives averaged 2,406 words (SD = 1,957; median = 1,855) and White participants' averaged 2,265 words (SD = 1,805; median = 1,706). Transcript length did not differ significantly across race, $t(892.2) = -1.13$, $p = .26$.

NEO-Personality Inventory-Revised (NEO-PI-R)

The NEO-PI-R (Costa & McCrae, 1992) is perhaps the most widely used measure of the FFM of personality and consists of 240 questions rated on a Likert-type scale from *strongly disagree* to *strongly agree*. Each domain (extraversion, agreeableness, conscientiousness, neuroticism, and openness) was scale scored. If a scale was missing one or two items, items that had been completed were averaged. If missing more than two items, these entries were deleted. The NEO-PI-R domain scores have shown strong internal consistency, test-retest reliability, and criterion validity in the SPAN dataset (J. R. Oltmanns et al., 2020; Wright et al., 2022).

Analysis

Complete details regarding transcription, transcript deidentification, and text preprocessing are provided in Oltmanns et al., (2026). Audio files were transcribed manually using Speechpad. Transcripts were deidentified using the "en_core_web_trf" model from the spaCy python package and preprocessed by removing transcript annotations and interviewer language.

Because matching reduced the White sample, we examined whether the findings generalize beyond the matched subsample. Following Rai et al. (2024), the primary analyses matched the White group to the Black group to test whether performance differences were attributable to sample-size imbalance or majority-group composition. Analyses using (a) the full White sample and (b) an alternatively matched reduced White sample constructed to resemble the full White sample were also conducted (See Supplement Tables).

Linguistic Inquiry and Word Count (LIWC)

Transcripts were processed through LIWC software (Boyd et al., 2022) to score them for psychological processes and parts of speech. Deidentified texts were entered and 117 variables were computed separately for the matched Black and White samples. Emojis were excluded. Each LIWC score was correlated with one personality trait one at a time for each race. Then, Fisher r-to-z transformation was used to identify significant differences in correlations between Black and White samples.

BERTopic

BERTopic (Grootendorst, 2022) can be used to identify frequent topics in life narrative interviews. It embeds sentences in documents (using Sentence BERT [sBERT]), reduces dimensionality, clusters the embeddings into topics, and tokenizes and weights the topics. Uniform Manifold Approximation and Projection (UMAP) reduces the dimensionality, and the

following settings were used: 15 nearest neighbors, 5 components, 0.0 min_dist, cosine metric. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) clustered the resulting components. All other settings were maintained at their default values. Finally, class-based Term Frequency-Inverse Document Frequency (cTF-IDF) tokenized the topics.

BERTopic automatically identifies one topic per document. However, participants discuss many different topics in their life narratives. We divided each transcript into utterances that were then fed into BERTopic, allowing each participant to be scored for multiple topics. Probabilities for each participant discussing each topic were calculated, and the maximum probability across all utterances for each participant, for each topic, was used as a single probability value. This max probability was then correlated with personality scale scores. This method assumes linear relationships between topics and personality traits. Each topic probability was correlated with one personality trait one at a time for each race. Then, Fisher r-to-z transformation was used to identify significant differences in correlations between the matched Black and White samples.

Fine-Tuning of RoBERTa

The RoBERTa-large model, an optimized version of Google's BERT language model (Devlin et al., 2019) by Facebook AI (Liu et al., 2019) was accessed using the simpletransformers library (Rajapakse, 2024). It was trained using the masked-language modeling objective on large, diverse text corpora. The model was fine-tuned on the life narrative interviews using a regression task with the following hyperparameters: a maximum sequence length of 512 tokens (required for RoBERTa), and batch sizes of 16 and 8 for training and

evaluation, respectively. A learning rate of 2e-5 was used, following best-performing configuration in Oltmanns et al. (2026).

RoBERTa-large was selected for two reasons. First, the present study was designed to evaluate an already-validated pipeline across Black and White older adults rather than to introduce a new model. RoBERTa was the best-performing model in Oltmanns et al. (2026), and holding the architecture fixed allows any group differences to be examined without confounding them with a change in architecture. Second, RoBERTa models are widely used for sentence- and document-level regression in psychological text analysis, and have demonstrated validity across personality and other clinical constructs (e.g., PTSD; Kjell et al., 2026), situating the present findings within a broader and comparable literature.

Given RoBERTa's 512-token limit, a sliding window approach divided the longer life-narrative texts into smaller chunks. Rather than truncating each interview to its first 512 tokens, embeddings were generated for each chunk and aggregated so that the full interview contributes to the final prediction. Specifically, we averaged the classification embeddings of all chunks from RoBERTa's last layer to create a single embedding per participant, which was then used to predict the final score with a feed-forward network. We note, however, that averaging is order-invariant and therefore captures local content more than global narrative structure, such as temporal sequencing across chapters and overall narrative coherence. Because personality is likely expressed not only through local linguistic content but also through the organization of the life story as a whole, segment-level prediction with averaging may underrepresent these global narrative features.

We used five-fold cross validation and three data splits – train, validation, and a held-out test set. Validation was a randomly selected 5% of the train set. After training, epochs were

tested on the validation set. The validation set results were used to determine the need for early stopping (i.e., if after five continuous epochs, performance did not improve). Trained model weights were then saved, and the model was put in evaluate mode and then evaluated on the hold-out test set. All results are reported based on the test set. Evaluation metrics were mean squared error (MSE), mean absolute error (MAE), and R-squared (R^2) score. Pearson r correlations between the RoBERTa-predicted and true scores are reported.

The RoBERTa model was fine-tuned using the Black, White, and combined samples as training sets subsequently for the regression task of predicting FFM traits in the Black and White test samples separately and together (in a combined sample test set). NEO-PI-R domain scores were standardized (z-scored) prior to modeling using scikit-learn's StandardScaler. The scaler was fit once on the full sample before cross-validation, and standardization was performed across the pooled sample. To assess cross-group performance, one model from each fold trained on the White sample was used to predict FFM traits in the Black sample, and vice versa. Metrics were then averaged across five folds.

Bias Analyses

A multi-component bias analysis was conducted to further evaluate potential racial bias in model predictions. Because no single fairness criterion is sufficient for regression models, and because several common fairness criteria are known to be mutually incompatible when group distributions differ (Chouldechova, 2017; Kleinberg et al., 2016), we operationalize fairness using several complementary criteria drawn from the algorithmic fairness literature (Barocas et al., 2019; Chouldechova, 2017). These map onto the components reported below: accuracy and rank-order correspondence (RMSE, R^2, r); error magnitude (Performance Bias: ΔMAE); raw mean prediction bias (Mean Bias); bias conditional on true score (Directional Bias); and

calibration (Calibration Analysis), with Distributional Diagnostics providing the variance context needed to interpret error-based metrics. Each criterion captures a different aspect of fairness, and the criteria can diverge in practice. Therefore, we interpret the findings as a profile across criteria rather than as a single fair/unfair verdict, and we do not claim to test all possible definitions of algorithmic fairness.

All analyses were implemented separately for each personality trait. Out-of-sample predictions from the held-out test sets of five folds were pooled for bias analyses, ensuring that all reported results reflect performance on unseen data.

Model performance between Black and White samples was compared using several metrics:

Accuracy and Rank-Order Correspondence. Root mean squared error (RMSE), coefficient of determination (R^2), and Pearson correlation (r) were computed for each racial group to evaluate if the model performs better for one group in terms of explained variance (R^2) and ranking individuals (r). To test whether rank-order accuracy differed across race rather than inferring it from the two separate correlations, we compared the predicted–true correlations for Black and White participants using a Fisher r-to-z test for independent correlations, with Cohen's q as the effect size.

Mean Bias. Mean prediction bias for each group was calculated as the mean of predicted score minus the true score (predicted - true). Positive results indicate overprediction while negative results indicate underprediction. To test whether mean bias differed across race, we compared the White and Black groups' signed errors using a Welch two-sample t-test, noting that mean bias does not adjust for group differences in true-score levels.

Distributional Diagnostics. Mean and variance of true scores and predicted scores within each group were calculated to contextualize accuracy metrics. Group differences in true-score variance were tested using Levene's test. These diagnostics evaluate whether observed performance differences could be driven by differences in score dispersion rather than model bias.

Because MAE and RMSE are expressed in the scale of the outcome, they can be influenced by group differences in true-score variance. A group with lower true-score variance may show lower absolute error even if the model captures less variance or ranks individuals less accurately within that group. For this reason, MAE and RMSE were interpreted alongside R^2, Pearson correlations, Levene tests of true-score variance, and calibration diagnostics. When true-score variance differed across groups, we placed greater interpretive emphasis on the full pattern across metrics rather than treating error magnitude alone as evidence of superior or inferior model performance.

Performance Bias. Absolute errors were compared to assess if prediction errors systematically differ in magnitude between racial groups. In other words, it tests whether the model is more or less accurate for one group, independent of error direction. Group differences were evaluated using observed difference in mean absolute error (ΔMAE) between Black and White samples, permutation tests to evaluate if ΔMAE differed by chance, and Cohen's *d* for ΔMAE as an effect-size measure.

Directional Bias. Error regression analysis was conducted to test whether the model systematically over or underpredicted for one racial group compared to the other, after controlling for true score. Prediction error was defined as predicted score minus true score. Pooled test-set predictions were fitted into the following linear regression model:

$Error = b_0 + b_1(race) + b_2(true\ score)$, where race was coded as 1 for Black participants and 0 for White participants.

b_1 represents directional bias. Specifically, $b_1 > 0$ indicates overprediction for Black participants relative to Whites and $b_1 < 0$ indicates underprediction for Black participants relative to Whites. Statistical significance of b_1 was assessed using two-tailed tests. Partial R^2 for b_1 was calculated to quantify the proportion of error variance uniquely explained by race.

Calibration Analysis. Calibration slope analysis was used to examine to what extent the variance in predicted scores align with the variance in true scores across the score range. A slope of 1 indicates perfect calibration. Slopes < 1 indicate overdispersion while slopes > 1 indicate underdispersion. t-tests were used to determine whether slopes deviated statistically significantly from 1. Calibration results were interpreted alongside variance diagnostics to assess whether differences in predictive spread reflected modeling artifacts or underlying distributional differences. To test whether calibration differed across race, rather than inferring a difference from separate per-group tests of whether each slope departed from 1.0, we fit a single model per trait regressing true scores on predicted scores, race (Black = 1, White = 0), and their interaction

$$true\ score = b_0 + b_1(predicted\ score) + b_2(race) + b_3(predicted \times race)$$

The predicted × race interaction (b_3) directly tests whether the calibration slope differs across groups.

Results

Descriptives for the NEO-PI-R scores are presented in Table 2 by race. The NEO-PI-R scores were continuous and approximately normally distributed.

Throughout, we report two complementary aggregations of model performance. In the predictive-performance results, R^2, r, and error metrics are computed within each cross-

validation fold and then averaged across the five folds (within each training/test condition). In the bias analyses, predictions from the held-out test sets of all five folds are first pooled into a single out-of-sample set, and metrics are then computed once on that pooled set. Therefore, pooled estimates can differ from the corresponding fold-averaged values.

LIWC

LIWC scores for the full sample are presented in Oltmanns et al., (2026). Using significance level of 0.01, there were 7 significant differences in the associations of LIWC scores and FFM scores between Black and White samples (Supplemental Table 1; absolute *z* ranged from 2.58 to 2.94). These differences were small in magnitude (Cohen's q ranged from .18 to .20), consistent with the exploratory nature of these analyses. With 117 LIWC variables tested across 5 traits (585 tests) at $\alpha = .01$, approximately six significant differences would be expected by chance alone, and we observed seven. Therefore, we treat these cross-group differences as hypothesis-generating rather than confirmatory and foreground effect sizes over statistical significance. These significantly different associations are displayed visually in Figure 2 (three from extraversion) and Figure 3 (one from neuroticism, agreeableness, conscientiousness, and openness each).

Descriptively, 50 personality-LIWC correlations reached significance among Black participants and 59 among White participants, with 16 significant in both samples. Given the large number of tests, we treat these counts as descriptive only and emphasize effect-size magnitude over the number of significant correlations. Effect sizes were small according to subjective guidelines (Cohen, 1992), with a median $r = .13$ in both groups. All correlations are displayed in Supplemental Table 2.

BERTopic

Topics for the full sample are presented in Oltmanns et al. (2026). There were 5 significant differences across race in topic correlations with personality traits at $p < .01$. These were extraversion with topics 139 (PhD dissertation) and 20 (teaching); conscientiousness with topic 97 (medication) and topic 0 (marriage), and neuroticism with topic 95 (communication style), with absolute value of z ranging from 2.61 to 3.37 and significant at $p < .01$ (Supplemental Table 3; Figure 4). These differences were small in magnitude (Cohen's q ranged from .17 to .22).

Overall, there were 10 significant correlations between topic probabilities and personality traits in the White sample, while there were 8 in the Black sample. All were within the absolute value effect sizes of $r = .12$ and $r = .18$ and statistically significant at $p < .01$. Among White participants, openness and conscientiousness were significantly associated with 2 topics each (42 [music] and 28 [life transitions] for the former, and 69 [reading aloud] and 104 [inner peace] for the latter). Also, among White participants, agreeableness and extraversion were significantly associated with one topic each (5 [military service] and 139 [PhD dissertation], respectively), and neuroticism was significantly correlated with four topics: 41 (depression), 89 [anger], 78 [travel], and 115 (school administration). Among the Black participants, openness, neuroticism, and extraversion were each significantly correlated with one topic (20 [teaching], 97 [medication], and 0 [marriage], respectively), conscientiousness was significantly correlated with two topics (0 [marriage] and 14 [crime]), and agreeableness was significantly correlated with three topics (6 [substance use], 21 [parents], and 14 [crime]).

Because the LIWC and BERTopic analyses involved many feature-level comparisons, these results should be interpreted as exploratory. We used a more conservative alpha level of .01 to reduce the likelihood of Type I error, but this threshold does not eliminate the possibility

of false-positive findings given the number of tests. Accordingly, we focus primarily on the magnitude and consistency of the observed associations rather than treating individual statistically significant LIWC categories or topics as definitive evidence of specific linguistic mechanisms.

RoBERTa

Generally, R^2 values showed better prediction from combined training sets, with the exceptions of extraversion and openness (Table 3). Openness showed some of the largest effect sizes—for both White and Black test sets—as well as extraversion, but only for the White test sets. Extraversion did not show predictive performance in the Black test sets as in the White test sets. Across both groups, RoBERTa models' averaged R^2 values across all training sets best predicted openness (e.g., White $R^2 = .20$, Black $R^2 = .13$). These sizes (with corresponding averaged Pearson *r* values of .48 and .40 for White and Black, respectively) were labeled subjectively as moderate by Cohen (1992) and large by Funder and Ozer (2019). Specifically in the context of predicting personality from language and characteristics of the test (e.g., it is a true multimethod effect), these effect sizes are considered large (see Oltmanns et al., 2026 discussion).

For White participants, RoBERTa predicted extraversion second best (averaged $R^2 = .10$, with R^2 of .13 in the White training-White test combination), but R^2 for Black participants for extraversion was only .01, and the White training-Black test combination (averaged $R^2 = .07$) was better than the Black training-Black test combination ($R^2 = .005$). This indicates the RoBERTa extraversion model worked best (second best overall) only for White participants and not for Black participants. Excluding extraversion, the next best performing predictive model was agreeableness (averaged $R^2 = .06$ for both Black and White), followed by neuroticism

(averaged R^2 = .05 for both Black and White). These two models did not differ significantly across test groups, indicating no difference in RoBERTa model performance by race across agreeableness and neuroticism. As in Oltmanns et al., (2026), the conscientiousness model showed the least predictive value, with a negative averaged R^2 value for the White test sets and .01 for the Black test sets. Averaged Pearson *r* values were .13 and .18 for White and Black, respectively. This indicates that the RoBERTa model worked least well for conscientiousness across both Black and White groups.

Bias Analyses

Neuroticism (Table 4)

Accuracy and Performance Bias. Pooled MAE was lower for the Black sample than the White sample, and this difference was statistically significant, although the effect size was small (Cohen's d = -.19). A similar pattern was observed for RMSE values. Mean prediction bias indicated small underprediction for White sample and overprediction for Black sample (pooled mean $\text{bias}_{\text{White}}$ = -.09; pooled mean $\text{bias}_{\text{Black}}$ =.06). Without controlling for true scores, mean prediction bias differed significantly across race ($p < .05$).

Despite the racial differences in error magnitude, model fit was similar across groups (R^2 = .07). Correlation between predicted and true neuroticism scores of each group was moderate (e.g., r = .28/.29) and statistically significant (both $p < .001$). The predicted-true correlation did not differ significantly across race for neuroticism (z = -.16, p = .88). The Black group had a significantly lower true-score variance compared to the White sample, indicated by a significant Levene test (p = .003).

Directional Bias. Race did not significantly predict error when controlling for true Neuroticism scores (b_1 = -.03, p = .17). The magnitude of this effect was almost zero (partial R^2

= .002), indicating no directional bias. In other words, the model did not systematically over- or under-predict one race against the other.

Calibration. Both groups showed calibration slopes below 1 (i.e., .85/.82), but neither was statistically significant, suggesting no over- or under-dispersion of predicted scores. Calibration-slope difference was not significant across race for neuroticism (p = .91).

Extraversion (Table 5)

Accuracy and Performance Bias. There were not statistically significant differences in absolute error between Black and White participants (p = .30, Cohen's d = -.06), nor in RMSE values. Mean prediction bias indicated small overprediction for the White sample and underprediction for the Black sample (pooled mean $\text{bias}_{\text{White}}$ = .04; pooled mean $\text{bias}_{\text{Black}}$ = -.07). Without controlling for true scores, mean prediction bias did not differ significantly across race (p = .083).

Although model fit was relatively weak and correlations between predicted and true scores was low but significant for both groups, the model predicted substantially better for the White participants than for the Black participants (pooled R^2_{White} = .14 and pooled r_{White} = .41; pooled R^2_{Black} = .01 and pooled r_{Black} = .17; both p <.001). The predicted-true correlation differed significantly across race for extraversion (z = 4.04, p <.001, Cohen's q = .27).

There were not statistically significant differences in true score variance, as indicated by the Levene test (p = .24). Although the pooled MAE and RMSE were slightly smaller for Black participants than for White participants, this did not appear to be driven by differences in true variance. True and predicted variances were relatively similar. However, model fit was weaker for Black participants, indicating that the model captured individual differences in extraversion worse in the Black group despite similar or slightly smaller MAE and RMSE.

Directional Bias. Race significantly predicted error when controlling for true extraversion scores (b_1 = .05, p = .005), suggesting overprediction for Black participants. The magnitude of this effect was small (partial R^2 = .003), indicating a significant but negligible effect.

Calibration. The calibration slope was significantly above 1 for the White group and significantly below 1 for the Black group (White slope = 1.63, p < .001; Black slope = .61, p = .03), suggesting overdispersion for the Black group and severe underdispersion for the White group. Calibration slope differed significantly across race for extraversion (b_3 = -1.02, p <.001). Thus, while directional bias was modest, there was evidence of significant calibration bias.

Openness (Table 6)

Accuracy and Performance Bias. The model showed smaller absolute errors for Black participants in terms of pooled MAE and RMSE. MAE was lower in the Black sample compared to the White sample, and this difference was statistically significant (p = .004) although the effect size was small (Cohen's d = -.20). Mean prediction bias indicated small underprediction for both groups, slightly larger for Black group (pooled mean $\text{bias}_{\text{White}}$ = -.02; pooled mean $\text{bias}_{\text{Black}}$ = -.06). Without controlling for true scores, mean prediction bias did not differ significantly across race (p = .56). Despite the racial differences in error magnitude, model fit and correlations between predicted and true scores were relatively comparable for both groups, slightly better for White participants and were both significant (pooled R^2_{White} = .20 and pooled r_{White} = .46; pooled R^2_{Black} = .16 and pooled r_{Black} = .41; both p <.001). The predicted-true correlation did not differ significantly across race for openness (z = 1.03, p = .30). However, the Black sample had a significantly lower true-score variance (.82) compared to the White sample (1.16), as indicated by a statistically significant Levene test (p = .004).

Directional Bias. Race significantly predicted error while controlling for true openness scores (b_1 = -.21, p < .001), indicating that the Black sample was underpredicted compared to the White sample. Race accounted for a significant amount of the variance in error (partial R^2 = .09). Considering the scale of the coefficient, it appears that the openness model moderately underpredicted openness scores for Black participants compared to Whites.

Calibration. The White group showed a significant calibration bias, with a slope of 1.26 (p = .02), while the Black group did not (slope = 1.08, p = .51). This indicates underdispersion for White group openness predictions—Predicted scores varied less than true scores for Whites. Calibration-slope difference was not significant across race for openness (p = .27). The Black sample showed lower absolute errors in MAE and RMSE, which reflected lower true-score variance. Thus, in addition to moderate evidence of directional bias for the Black group, there was also evidence of calibration bias for the White group.

Agreeableness (Table 7)

Accuracy and Performance Bias. There were not statistically significant absolute errors across groups. Mean prediction bias indicated small underprediction for Black participants and overprediction for Whites, slightly larger bias for Black group (pooled mean $bias_{White}$ =.01; pooled mean $bias_{Black}$ = -.02). Without controlling for true scores, mean prediction bias did not differ significantly across race (p = .70).

Model fit was the same for both groups (R^2 = .06). Correlations between predicted and true agreeableness scores were relatively similar and were both significant (pooled r_{White} = .26 and pooled r_{Black} = .28; both p <.001). The predicted-true correlation did not differ significantly across race for agreeableness (z = -.25, p = .80). These results indicate no racial differences in

accuracy. True-score variances were similar across groups (1.00), supported by Levene test (p = .70).

Directional Bias. Race significantly predicted error when controlling for true agreeableness scores (b_1 = -.08, p = .002), suggesting underprediction for Black participants compared to Whites. The magnitude of this effect was relatively low (partial R^2 = .004), indicating a significant but negligible effect.

Calibration. Both groups showed calibration slopes below 1, and both were statistically significant, (White group slope = .72, p = .02; Black group slope = .72, p = .02), suggesting overdispersion for both groups. Calibration-slope difference was not significant across race for agreeableness (p = .99).

Conscientiousness (Table 8)

Accuracy and Performance Bias. The model did not show statistically different absolute errors across groups. Mean prediction bias indicated small underprediction for Black and overprediction for White group (pooled mean $bias_{White}$ = .15; pooled mean $bias_{Black}$ = -.13). Without controlling for true scores, mean prediction bias differed significantly across race (p < .001). Model fit was either small or poor in both groups, with small and negative R^2 values. This dovetails with the findings of Oltmanns et al., (2026). Correlations between predicted and true conscientiousness scores of both groups were small but significant (pooled r_{White} = .14, pooled r_{Black} = .19, and both p <.001). The predicted-true correlation did not differ significantly across race for conscientiousness (z = -.80, p = .42).

True-score variance did not significantly differ between groups, supported by Levene test (p = .55).

Directional Bias. Race significantly predicted error when controlling for true conscientiousness scores (b_1 =.13, $p < .001$), suggesting overprediction for Black participants compared to Whites. The magnitude of this effect was small (partial R^2 = .01), indicating small directional bias.

Calibration. The White group showed a significant calibration bias, with a slope of 0.53 (p = .01), while the Black group did not (slope = .84, p =.43). This indicates overdispersion for the White group conscientiousness predictions—Predicted scores varied more than true scores for Whites. Calibration-slope difference was not significant across race for conscientiousness (p = .26). Although statistically significant directional bias was observed, the model showed little predictive utility overall, limiting the practical significance of these findings.

To assess whether these patterns held beyond the matched subsample, we repeated the RoBERTa analyses using (a) the full White sample and (b) an alternatively matched White sample constructed to resemble the full White sample (Supplemental Tables 5–14). These supplemental analyses generally supported the primary findings. Specifically, the extraversion performance gap and the openness directional bias underpredicting scores for Black participants were present in all three samples. The extraversion gap was somewhat attenuated, and the openness directional bias was larger in the full White sample (partial R^2 = .09 in the matched sample vs. .15 and .12 in the full and alternatively matched samples), indicating the central findings were not specific to the matched subsample.

Discussion

Advances in AI have exciting implications for psychological assessment. In particular, language-based methods could make multimethod personality assessment more scalable by scoring behavior people already generate. However, little-to-no research has examined whether

language models of personality operate equally across racial groups. Prior studies indicate that language models of depression may operate differently across Black and White American adults. This study indicates that certain language models of FFM personality may also operate differently across Black and White older adults.

We first consider predictive performance in the test data for each group. The predicted–true correlations did not differ significantly across race for neuroticism, openness, agreeableness, or conscientiousness (all $p > .30$). Extraversion was the exception: predictions correlated with true scores at $r = .42$ for White participants but only $r = .17$ for Black participants, a significant difference ($z = 4.04$, $p < .001$, Cohen's q = .27) and the only significant racial difference in predictive accuracy across the five traits. For context: because the model predicts self-reported personality from an entirely independent source—narrative language that never asked about personality—correlations near $r = .40$ represent substantial multimethod effects, approaching the range considered strong for same-method convergent validity (Clark & Watson, 2019; Funder & Ozer, 2019). The fairness question here, however, concerns not their absolute magnitude but their comparability across groups.

Error-based metrics (RMSE and MAE) showed modest racial differences for neuroticism and openness, with the Black sample showing slightly smaller absolute errors than the White sample. Lower absolute error did not indicate better prediction for the Black group, however, because the reduced errors coincided with significantly lower true-score variance in the Black group, and MAE and RMSE are sensitive to the dispersion of the outcome variable. Reduced criterion variance constrains the possible range of prediction errors and can lower absolute error even when the model captures less of the group's true-score variance. Meaningful cross-group comparison of predictive accuracy therefore requires reasonably comparable criterion variance.

True-score variance was comparable across groups for extraversion, agreeableness, and conscientiousness, but significantly lower in the Black group for openness and neuroticism (Tables 4-8)—precisely the traits where the Black group's absolute errors were smaller. This pattern suggest that the reduced error magnitude reflects distributional properties rather than better predictive ability. Because R^2 and Pearson r are scale-free, they provide a more defensible basis than MAE or RMSE for comparing predictive performance across groups whose criterion variances differ. Therefore, we foreground R^2 and r in our cross-group conclusions and interpret MAE and RMSE jointly with the Levene tests and calibration results.

In addition to predictive performance, it is also important to consider mean-level model predictions across groups. Directional bias was statistically significant for every trait except neuroticism, but only openness showed a meaningful effect: the model underpredicted openness for Black participants after controlling for true scores (b = -.21, partial R^2 = .09); the remaining effects were negligible (partial $R^2 \leq .01$). Calibration also differed across race, though formal testing showed this was reliable only for extraversion: the predicted x race interaction was significant ($b = -1.02$, $p < .001$), with predictions underdispersed relative to true scores for White participants (slope = 1.63) and overdispersed for Black participants (slope = 0.61). Openness showed a numerically similar but nonsignificant asymmetry (interaction p = .27), so we treat its mean-level directional bias as the substantive openness finding. For extraversion, then, comparable average error coexisted with different relationships between predicted and true scores across groups. The conscientiousness model showed the poorest predictive performance in both groups (consistent with Oltmanns et al., 2026). Because its predictive validity was low, we do not interpret its directional-bias or calibration patterns; structure in predictions largely unrelated to true scores should not be read as evidence of bias. We instead treat the poor

performance itself as the finding—and as an illustration of the broader point that predictive validity must be established before subgroup bias metrics can be interpreted.

Two non-exclusive explanations count account for the extraversion finding. First, RoBERTa may represent linguistic patterns unevenly across groups because of its pretraining data, tokenization, or the present processing pipeline. Second, the NEO-PI-R criterion may not align equally with the linguistic expression of extraversion across groups—a difference in how the construct is measured rather than in model quality. Tay et al. (2022) describe this as ground-truth bias: when the criterion functions non-equivalently across groups, a model trained to predict it is optimized against a differentially valid target, and they accordingly recommend establishing measurement invariance of the criterion before interpreting downstream model bias. Although the NEO-PI-R has shown strong reliability and criterion validity in the SPAN sample overall (J. R. Oltmanns et al., 2020; Wright et al., 2022), its measurement invariance across Black and White Americans has not, to our knowledge, been directly established—a gap that extends to FFM self-report inventories more generally. Where FFM invariance has been tested across ethnoracial groups, it has typically been only partial, with extraversion among the least invariant domains (Lui et al., 2020; Rollock & Lui, 2016), and evidence for a related maladaptive personality trait measure across Black and White Americans is mixed (Bagby et al., 2023; Freilich et al., 2023). That extraversion showed the largest performance gap here is consistent with a criterion-related contribution, though the present data cannot distinguish it from model-related differences. Testing invariance directly via multi-group confirmatory factor analysis or differential item functioning (Holland & Thayer, 1988) is an important next step—and our recommendation that measures be validated on the groups with which they are used applies to the self-report criterion as much as to the language model.

The exploratory LIWC and BERTopic analyses did not identify a clear mechanism for the extraversion performance gap, but they offer descriptive context on where cross-group differences appear—useful given that the RoBERTa model itself is a black box. Extraversion, the trait with the largest accuracy gap, also showed several cross-group feature differences. Extraversion was significantly more positively related to discussion of work and lifestyle activities for Black participants but not White participants, negatively related to LIWC dictionary word use for Black participants but not White participants, and more positively related to talking about teaching and graduate schoolwork for Black participants. These differences converge on a common theme of work and lifestyle. Given the number of comparisons, we interpret them cautiously and do not treat them as a confirmed explanation for the extraversion gap, which may instead reflect differential linguistic expression of extraversion across groups and/or a mismatch with the self-report criterion, in addition to model-related factors.

Several other LIWC differences emerged across groups. As noted in the Results, the seven significant LIWC differences were close to the number expected by chance across this many tests, so the interpretations below are offered as hypotheses for future work rather than established patterns—they should not be read as evidence that the personality constructs carry different meanings across groups, only that some linguistic indicators may differ in this sample. LIWC Risk words (e.g., secur*, protect*, pain, risk*) were associated with neuroticism for White but not Black participants; one possibility is that such language reflects structural conditions (e.g., neighborhood safety, economic precarity) more than internal disposition for Black older adults in this cohort (Williams & Mohammed, 2009), making it a less diagnostic marker of trait neuroticism. LIWC Anxiety words (worry, fear, afraid) were more strongly associated with conscientiousness for Black than White participants, consistent with the John Henryism and

skin-deep-resilience literatures, in which high-effort coping under structural adversity co-occurs with anxiety and vigilance (Brody et al., 2013; James, 1994; Miller et al., 2015). And LIWC Memory words (remember, forget, remind) were more strongly associated with agreeableness for Black than White participants: in autobiographical and oral traditions, recollecting family and communal history is itself a relational act (Smitherman, 1977), which would make memory language a more agreeableness-relevant marker for Black participants. Given the number of comparisons, we do not treat any single association as definitive.

Limitations and Future Directions

The observed differences should not be interpreted solely as evidence about language–personality associations; model- and pipeline-related factors may also contribute, and the present design cannot separate them. RoBERTa was pretrained on corpora—English Wikipedia, BookCorpus, CC-News, Stories, and Reddit-sourced web text—whose contributors skew toward younger and already-privileged users, so the language practices of marginalized communities are underrepresented relative to their prevalence in the population (Bender et al., 2021). Independently, NLP systems perform less accurately on text containing features associated with African American English (Blodgett et al., 2020). The resulting representations may therefore encode some linguistic and topical content more reliably than others. In addition, the 512-token input limit required segmenting each interview, and averaging chunk embeddings is order-invariant, so global narrative structure—temporal sequencing, coherence across chapters—may be underrepresented, potentially asymmetrically if narrative organization differs across groups.

These findings apply to the RoBERTa pipeline used here rather than to language-based personality assessment generally. Alternative representations would address some contributing factors but not others. Simpler word-level approaches such as TF-IDF avoid subword

fragmentation and do not rely on a pretrained corpus, but they typically predict less accurately and would relocate rather than remove the representation problem. Other transformer variants such as DeBERTa share RoBERTa's subword tokenization and predominantly English pretraining and would be unlikely to differ substantially. Long-context or sequence-preserving architectures could address the segmentation and averaging constraints, although early evidence on these narratives is mixed: Hussain et al. (2026) found that long-context models did not outperform segment-based approaches on this dataset, though an architecture that preserved sequence across windows did improve prediction. Cross-architecture comparison is therefore warranted, but no change in model would address the self-report criterion or the culturally patterned ways people tell their life stories. Finally, transcript length was comparable across groups and is unlikely to account for the observed differences, consistent with Hussain et al. (2026), who found transcript length unrelated to prediction performance on these interviews.

The original sample contained a larger proportion of White than Black participants, consistent with the demographics of the St. Louis area from which it was drawn, so matching produced equally sized groups but reduced the analytic sample to 450 participants per group, limiting power and the precision of subgroup estimates. Prior work suggests effect sizes in language-based prediction begin to stabilize at a few thousand participants (Eichstaedt et al., 2020), so larger samples are needed—especially to examine within-group heterogeneity, adjust for additional covariates, and test intersectional differences. Additionally, standardization parameters were estimated on the pooled sample prior to cross-validation. This is a minor source of information leakage that is unlikely to affect cross-group comparisons.

Relatedly, our analyses treat Black and White participants as broad social groups, although each encompasses substantial variation in region, socioeconomic position, education,

dialect use, immigration history, and cultural identity, and the present sample did not support reliable analysis of this heterogeneity. Race should therefore be understood here as a coarse social classification rather than a homogeneous category. Matching on education and income reduced the possibility that observed differences reflect correlated socioeconomic factors rather than race per se, but it did not eliminate other correlated differences, and we lacked power to test whether the observed patterns vary by gender or other intersecting characteristics. Future work with larger, more diverse samples should examine these possibilities directly.

Future work should examine whether model scores show comparable construct validity including convergent validity with informant reports and criterion validity for consequential life outcomes across racial groups (Oltmanns et al., 2026). Direct tests of NEO-PI-R measurement invariance would also help distinguish criterion-related from model-related explanations. Similar fairness analyses should be extended to other constructs that can be assessed from life narratives, including depression, anxiety, and self-esteem (Rai et al., 2024).

Generalizability

These results apply to Black and White English-speaking older adults recruited from a single U.S. metropolitan area, assessed with the NEO-PI-R and a 20-minute life narrative interview. The matched analytic sample was constructed to balance education and income, so it is not representative of the broader White population in this cohort. Whether these findings extend to other racial and ethnic groups, younger cohorts, other regions, other personality inventories, or other narrative formats is unknown.

Implications

These findings carry concrete implications for applied use of language-based personality models. In settings where such models might inform consequential decisions such as clinical

screening, personnel selection, or large-scale research, differential performance across racial groups could translate into unequal accuracy or systematic over- or under-prediction for minoritized groups, with real consequences. Our results suggest that aggregate predictive accuracy is an insufficient basis for deployment. Specifically, a model can perform well overall while performing markedly worse, or showing directional bias, for a subgroup (as with extraversion here). Practically, this argues for routine subgroup evaluation (accuracy, directional bias, and calibration assessed separately by group) before any applied use, and for training researchers and clinicians who encounter AI-based assessment tools to interpret these subgroup metrics and to recognize the ethical limits of deploying models in contexts where differential performance could shape opportunities or outcomes.

Conclusions

Models using AI for psychological assessment need to be validated in the populations in which they will be used. The present study is the first, to our knowledge, to examine the performance of language-based AI assessment of personality across Black and White Americans. Predictive performance was broadly similar for most traits, but two differences emerged: the extraversion model predicted far less accurately for Black than White participants, and calibration differed across groups, and the openness model underpredicted mean levels for Black participants after controlling for true scores. These findings are initial evidence that researchers should evaluate racial group differences in language-based AI models of personality before such models are applied.

References

Aguirre, C. A., Harrigian, K., & Dredze, M. (2021, April). Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2932-2949).

Allport, G. W., & Odbert, H. S. (1936). Trait-Names. A Psycho-lexical Study. *Psychological Monographs*, *47*(1), i–171. https://doi.org/10.1037/h0093360

APA Task Force on Psychological Assessment and Evaluation Guidelines. (2020). *APA Guidelines for Psychological Assessment and Evaluation* (Nos. 510142020–001). American Psychological Association. https://doi.org/10.1037/e510142020-001

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Bagby, R. M., Keeley, J. W., Williams, C. C., Mortezaei, A., Ryder, A. G., & Sellbom, M. (2022). Evaluating the measurement invariance of the Personality Inventory for DSM-5 (PID-5) in Black Americans and White Americans. *Psychological Assessment*, *34*, 82-90.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. https://fairmlbook.org

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer (arXiv:2004.05150). arXiv. https://doi.org/10.48550/arXiv.2004.05150

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). 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. https://doi.org/10.1145/3442188.3445922

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Boyd, R. L., Ashwini Ashokkumar, Seraj, S., & Pennebaker, J. W. (2022). *The Development and Psychometric Properties of LIWC-22*. https://doi.org/10.13140/RG.2.2.23890.43205

Brickman, J., Gupta, M., & Oltmanns, J. R. (2025). Large Language Models for Psychological Assessment: A Comprehensive Overview. *Advances in Methods and Practices in Psychological Science*, *8*(3), 1–26. https://doi.org/10.1177/251524592513435

Brody, G. H., Yu, T., Chen, E., Miller, G. E., Kogan, S. M., & Beach, S. R. H. (2013). Is Resilience Only Skin Deep? Rural African Americans' Preadolescent Socioeconomic Status-Related Risk and Competence and Age 19 Psychological Adjustment and Allostatic Load. *Psychological Science*, *24*(7), 1285–1293. https://doi.org/10.1177/0956797612471954

Chapman, B. P., Roberts, B., & Duberstein, P. (2011). Personality and longevity: Knowns, unknowns, and implications for public health and personalized medicine. *Journal of Aging Research*, 2011, 759170. https://doi.org/10.4061/2011/759170

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092–1122. https://doi.org/10.1037/a0021212

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*(1), 5–13.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv. http://arxiv.org/abs/1810.04805

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., ... & Ungar, L. H. (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*, 398-427.

Ferraro, K. F., & Shippee, T. P. (2009). Aging and cumulative inequality: how does inequality get under the skin? *The Gerontologist, 49*(3), 333–343. https://doi.org/10.1093/geront/gnp034

Freilich, C. D., Palumbo, I. M., Latzman, R. D., & Krueger, R. F. (2023). Assessing the measurement invariance of the Personality Inventory for DSM-5 across Black and White americans. *Psychological Assessment*, *35*, 721-728.

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Galton, F. (1884). Measurement of character. *Fortnightly Review*, *36*(212), 179–185.

Goldberg, L. R. (1993). The Structure of Phenotypic Personality Traits. *American Psychologist*, 9.

Green, L. J. (2002). *African American English: A linguistic introduction*. Cambridge University Press.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (No. arXiv:2203.05794). arXiv. http://arxiv.org/abs/2203.05794

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology, 107*(8), 1323–1351. https://doi.org/10.1037/apl0000695

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & I. Braun (Eds.), *Test validity* (pp. 129–145). Erlbaum.

Hussain, R., Ma, Z., Khandelwal, R., Oltmanns, J., & Gupta, M. (2026). Language-based personality assessment from life narratives: a focus on model interpretability and efficiency. *Frontiers in Artificial Intelligence*, *9*, 1760246.

James, S. A. (1994). John Henryism and the health of African-Americans. Culture, Medicine and Psychiatry, 18(2), 163–182. https://doi.org/10.1007/BF01379448

John, O. P. (2021). History, Measurement, and Conceptual Elaboration of the Big-Five Trait Taxonomy: The Paradigm Matures. In O. P. John & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (Fourth edition, pp. 35). The Guilford Press.

Kish, L. (1949). A Procedure for Objective Respondent Selection within the Household. *Journal of the American Statistical Association*, *44*(247), 380–387. https://doi.org/10.1080/01621459.1949.10483314

Kjell, O. N. E., Ganesan, A. V., Boyd, R., Oltmanns, J. R., Rivero, A., Feltman, S., Carr, M.A., Luft, B. J., Kotov, R., & Schwartz, H. A. (2026). Replicability and validity of a new AI assessment of PTSD from patient language: A sequential evaluation with model pre-registration. *Clinical Psychological Science.* https://doi.org/10.1177/21677026261439026/

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond Rating Scales: With Targeted Evaluation, Language Models are Poised for Psychological Assessment. *Psychiatry Research*, 115667. https://doi.org/10.1016/j.psychres.2023.115667

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores* (arXiv:1609.05807). arXiv. https://doi.org/10.48550/arXiv.1609.05807

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* arXiv. https://arxiv.org/abs/1907.11692

Lui, P. P., Samuel, D. B., Rollock, D., Leong, F. T., & Chang, E. C. (2020). Measurement invariance of the five factor model of personality: Facet-level analyses among Euro and Asian Americans. *Assessment*, *27*, 887-902.

McAdams, D. P. (1993). The Stories We Live by: Personal Myths and the Making of the Self. Guilford Press.

McAdams, D. P. (2001). The Psychology of Life Stories. *Review of General Psychology*, *5*(2), 100–122. https://doi.org/10.1037/1089-2680.5.2.100

McAdams, D. P., & Pals, J. L. (2006). A new Big Five: Fundamental principles for an integrative science of personality. *American Psychologist, 61*(3), 204–217. https://doi.org/10.1037/0003-066X.61.3.204

Miller, G. E., Yu, T., Chen, E., & Brody, G. H. (2015). Self-control forecasts better psychosocial outcomes but faster epigenetic aging in low-SES youth. *Proceedings of the National Academy of Sciences, 112*(33), 10325–10330. https://doi.org/10.1073/pnas.1505063112

Mroczek, D. K., & Spiro, A. (2007). Personality Change Influences Mortality in Older Men. *Psychological Science*, *18*(5), 371–376. https://doi.org/10.1111/j.1467-9280.2007.01907.x

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Oltmanns, J. R., Jackson, J. J., & Oltmanns, T. F. (2020). Personality change: Longitudinal self-other agreement and convergence with retrospective-reports. *Journal of Personality and Social Psychology*, *118*(5), 1065–1079. https://doi.org/10.1037/pspp0000238

Oltmanns, J. R., Khandelwal, R., Ma, J., Brickman, J., Do, T., Hussain, R., & Gupta, M. (2026). Language-based AI modeling of personality traits and pathology from life narrative interviews. *Journal of Psychopathology and Clinical Science, 135,* 122-135.. https://doi.org/10.1037/abn0001047

Oltmanns, T. F., Rodrigues, M. M., Weinstein, Y., & Gleason, M. E. J. (2014). Prevalence of Personality Disorders at Midlife in a Community Sample: Disorders and Symptoms Reflected in Interview, Self, and Informant Reports. *Journal of Psychopathology and Behavioral Assessment*, *36*(2), 177–188. https://doi.org/10.1007/s10862-013-9389-7

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In *Handbook of research methods in personality psychology* (pp. 224–239).

Peters, H., & Matz, S. C. (2024). Large language models can infer psychological dispositions of social media users. *PNAS Nexus, 3*(6), pgae231. https://doi.org/10.1093/pnasnexus/pgae231

Rai, S., Stade, E. C., Giorgi, S., Francisco, A., Ungar, L. H., Curtis, B., & Guntuku, S. C. (2024). Key language markers of depression on social media depend on race. *Proceedings of the National Academy of Sciences*, *121*(14), e2319837121. https://doi.org/10.1073/pnas.2319837121

Rajapakse, T. C. (2024). *Simple Transformers* [Computer software]. https://simpletransformers.ai/

Rickford, J. R. (1999). *African American vernacular English: Features, evolution, educational implications*. Malden, Mass. : Blackwell Publishers. http://archive.org/details/africanamericanv0000rick

Roberts, B. W., & Wood, D. (2014). Personality development in the context of the neo-socioanalytic model of personality. In Handbook of Personality Development (pp. 11–39). Taylor and Francis. https://doi.org/10.4324/9781315805610-9

Rollock, D., & Lui, P. P. (2016). Measurement invariance and the five-factor model of personality: Asian international and Euro American cultural groups. *Assessment*, *23*, 571-587.

Smitherman, G. (1977). Talkin and Testifyin: The Language of Black America. Boston: Houghton Mifflin.

Spence, C. T., & Oltmanns, T. F. (2011). Recruitment of African American men: Overcoming challenges for an epidemiological study of personality and health. *Cultural Diversity and Ethnic Minority Psychology*, *17*(4), 377–380. https://doi.org/10.1037/a0024732

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences, 117*(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. *Advances in Methods and Practices in Psychological Science*, *5*(1), 25152459211061337. https://doi.org/10.1177/25152459211061337

Terracciano, A., Löckenhoff, C. E., Zonderman, A. B., Ferrucci, L., & Costa, P. T. (2008). Personality predictors of longevity: Activity, emotional stability, and conscientiousness. *Psychosomatic Medicine*, *70*(6), 621–627. https://doi.org/10.1097/PSY.0b013e31817b9371

Turner, A. F., Couch, N. G., Cowan, H. R., Otto-Meyer, R., Murthy, P., Logan, R. L., ... & McAdams, D. P. (2022). The good and the bad in black and white: Stories of life's high and low points told by black and white midlife adults in America. *Journal of Research in Personality*, *101*, 104298.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300. https://doi.org/10.1037/a0017908

Williams, D. R., & Mohammed, S. A. (2009). Discrimination and racial disparities in health: Evidence and needed research. *Journal of Behavioral Medicine*, *32*(1), 20–47. https://doi.org/10.1007/s10865-008-9185-0

Wright, A. J., Weston, S. J., Norton, S., Voss, M., Bogdan, R., Oltmanns, T. F., & Jackson, J. J. (2022). Prospective self- and informant-personality associations with inflammation, health behaviors, and health indicators. *Health Psychology*, *41*(2), 121–133. https://doi.org/10.1037/hea0001162

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for Longer Sequences. *Advances in Neural Information Processing Systems*, *33*, 17283–17297. https://proceedings.neurips.cc/paper_files/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html

Table 1. *Participant Demographics*

Variable	N_{White}	N_{Black}	$\%_{White}$	$\%_{Black}$
	450	450		
Gender				
Female	262	256	58.4%	56.9%
Male	188	194	41.6%	43.1%
Marital Status				
Married	192	159	42.8%	35.3%
Divorced	136	139	30.1%	30.9%
Never Married	61	63	13.6%	14.0%
Widowed	31	48	6.9%	10.7%
Separated	7	20	1.6%	4.4%
Living with Partner	14	14	3.1%	3.1%
Other Serious Relationship	8	4	1.8%	0.9%
Education				
Some College	116	112	25.8%	24.9%
Bachelor Degree	103	67	22.9%	14.9%
Master Degree	44	40	9.8%	8.9%
Associate Degree	56	59	12.5%	13.1%
Doctorate	5	5	1.1%	1.1%
Vocational Tech Degree	31	44	6.9%	9.8%
Elementary or Junior High	7	12	1.6%	2.7%
GED	84	102	18.5%	22.7%
Doctorate	5	5	1.1%	1.1%
Professional Degree	2	2	0.4%	0.4%
Annual Household Income				
Under $20,000	62	95%	13.8%	21.1%
$20,000-$39,999	110	117%	24.5%	26%
$40,000-$59,999	112	105%	24.9%	23.3%
$60,000-$79,999	66	51%	14.7%	11.3%
$80,000-$99,999	45	29%	10.0%	6.4%
$100,000-$119,999	29	23%	6.5%	5.1%
$120,000-$139,999	6	6%	1.3%	1.3%
$140,000 or more	5	5%	1.1%	1.1%
Missing	15	19%	3.1%	4.2%

Table 2

Descriptive Statistics for the NEO-PI-R Personality Scores

Scaled FFM Score	M_W	M_B	SD_W	SD_B	$Skew_W$	$Skew_B$	Min_W	Min_B	Max_W	Max_B	$Median_W$	$Median_B$
Neuroticism	76.50	72.34	22.38	19.31	0.55	0.45	13	19	157	142.8	74	71
Extraversion	106.91	110.00	18.79	16.98	-0.17	-0.18	47	53	171	174	108	110
Openness	113.47	109.68	19.96	16.71	0.45	0.78	51	67	202.5	201	113	108
Agreeableness	131.51	130.58	16.54	16.51	0.34	0.17	85	66	210	192	131.5	130
Conscientiousness	119.76	127.61	17.59	18.90	-0.11	1.33	61	77	172	234	121	126

Table 3

RoBERTa Language Modeling of Personality

	White test set			Black test set			Combined test set		
Neuroticism	*MSE*	*r*	R^2	*MSE*	*r*	R^2	*MSE*	*r*	R^2
White training set	0.93	0.29	0.05	0.72	0.35	0.04			
Black training set	1.35	0.36	-0.01	0.95	0.27	0.05			
Combined training set	1.03	0.31	0.07	0.79	0.30	0.06	0.91	0.31	0.09
Extraversion	*MSE*	*r*	R^2	*MSE*	*r*	R^2	*MSE*	*r*	R^2
White training set	0.85	0.42	0.13	0.76	0.29	0.07			
Black training set	1.22	0.26	0.01	0.98	0.17	0.005			
Combined training set	0.96	0.41	0.1	0.90	0.24	-0.002	0.93	0.34	0.07
Openness	*MSE*	*r*	R^2	*MSE*	*r*	R^2	*MSE*	*r*	R^2
White training set	0.78	0.48	0.20	0.60	0.40	0.15			
Black training set	1.20	0.44	0.16	0.86	0.39	0.13			
Combined training set	0.89	0.49	0.23	0.81	0.40	0.12	0.81	.45	0.20
Agreeableness	*MSE*	*r*	R^2	*MSE*	*r*	R^2	*MSE*	*r*	R^2
White training set	0.96	0.26	0.02	0.92	0.29	0.07			
Black training set	0.89	0.35	0.11	0.97	0.25	0.03			
Combined training set	0.91	0.30	0.06	0.93	0.27	0.06	0.92	0.29	0.07
Conscientiousness	*MSE*	*r*	R^2	*MSE*	*r*	R^2	*MSE*	*r*	R^2
White training set	1.02	0.12	-0.03	1.33	0.18	-0.15			
Black training set	0.99	0.16	-0.15	0.98	0.18	0.02			
Combined training set	0.92	0.13	-0.05	1.02	0.22	0.01	0.97	0.23	0.03

Note . Neuroticism N_{White} = 445, N_{Black} = 444; Extraversion N_{White} = 446, N_{Black} = 443; Openness N_{White} = 443, N_{Black} = 442; Agreeableness N_{White} = 444, N_{Black} = 444; Conscientiousness N_{White} = 443, N_{Black} = 444.

Table 4

Bias Analyses of Neuroticism Model

Accuracy	*RMSE*	R^2	*r*	*Mean Bias*	$Mean_{true}$	$Mean_{pred}$	Var_{true}	Var_{pred}	*F*	*df1*	*df2*	*p*
White	1.03	0.07	0.28	-0.09	0.10	0.01	1.14	0.13	8.61	1	889	0.003
Black	0.88	0.08	0.29	0.06	-0.10	-0.04	0.84	0.11				
Performance Bias	*MAE*	*ΔMAE*	*p*	*Cohen's d*								
White	0.80											
Black	0.68	-0.11	0.004	-0.19								
Directional Bias	b_1	*p*	R^2									
	-0.03	0.17	0.002									
Calibration	Slope	*p*										
White	0.85	0.26										
Black	0.82	0.17										

Table 5

Bias Analyses of Extraversion Model

Accuracy	*RMSE*	R^2	*r*	*Mean Bias*	$Mean_{true}$	$Mean_{pred}$	Var_{true}	Var_{pred}	*F*	*df1*	*df2*	*p*
White	0.97	0.14	0.41	0.04	-0.09	-0.05	1.09	0.07	1.36	1	889	0.24
Black	0.94	0.01	0.17	-0.07	0.09	0.01	0.89	0.07				
Perform-ance Bias	*MAE*	*ΔMAE*	*p*	*Cohen's d*								
White	0.77	-0.05	0.3	-0.06								
Black	0.72											
Directional Bias	b_1	*p*	R^2									
	0.05	0.005	0.003									
Calibration	Slope	*p*										
White	1.63	<0.01										
Black	0.61	0.03										

Table 6

Bias Analyses of Openness Model

Accuracy	*RMSE*	R^2	*r*	*Mean Bias*	$Mean_{true}$	$Mean_{pred}$	Var_{true}	Var_{pred}	*F*	*df1*	*df2*	*p*
White	0.96	0.20	0.46	-0.02	0.10	0.08	1.16	0.16	8.17	1	885	0.004
Black	0.83	0.16	0.41	-0.06	-0.10	-0.16	0.82	0.12				
Performance Bias	*MAE*	*ΔMAE*	*p*	*Cohen's d*								
White	0.74	-0.11	0.004	-0.20								
Black	0.62											
Directional Bias	b_1	*p*	R^2									
	-0.21	0.00	0.09									
Calibration	Slope	*p*										
White	1.26	0.02										
Black	1.08	0.51										

Table 7

Bias Analyses of Agreeableness Model

Accuracy	*RMSE*	R^2	*r*	*Mean Bias*	$Mean_{true}$	$Mean_{pred}$	Var_{true}	Var_{pred}	*F*	*df1*	*df2*	*p*
White	0.97	0.06	0.26	0.01	0.03	0.03	1	0.13	0.15	1	888	0.70
Black	0.97	0.06	0.28	-0.02	-0.03	-0.05	1	0.15				
Performance Bias	*MAE*	*ΔMAE*	*p*	*Cohen's d*								
White	0.74	0.02	0.74	0.02								
Black	0.76											
Directional Bias	b_1	*p*	R^2									
	-0.08	0.002	0.004									
Calibration	Slope	*p*										
White	0.72	0.02										
Black	0.72	0.02										

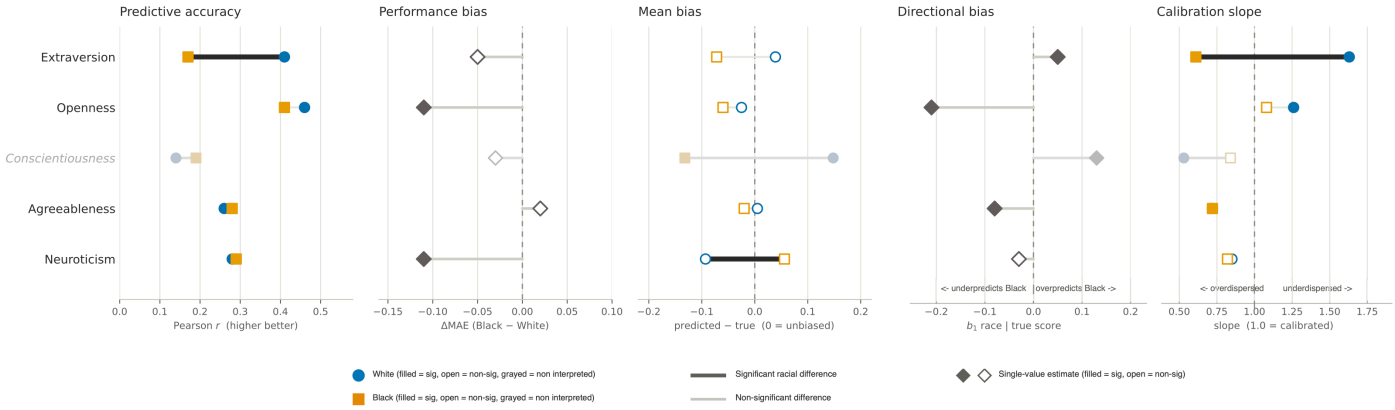
Table 8

Bias Analyses of Conscientiousness Model

Accuracy	*RMSE*	R^2	*r*	*Mean Bias*	$Mean_{true}$	$Mean_{pred}$	Var_{true}	Var_{pred}	*F*	*df1*	*df2*	*p*
White	0.95	-0.02	0.14	0.15	-0.21	-0.06	0.89	0.06	0.36	1	887	0.55
Black	1.00	0.02	0.19	-0.13	0.21	0.08	1.02	0.05				
Performance Bias	*MAE*	*ΔMAE*	*p*	*Cohen'd*								
White	0.75	-0.03	0.49	-0.05								
Black	0.72											
Directional Bias	b_1	*p*	R^2									
	0.13	<0.01	0.01									
Calibration	Slope	*p*										
White	0.53	0.01										
Black	0.84	0.43										

Figure 1

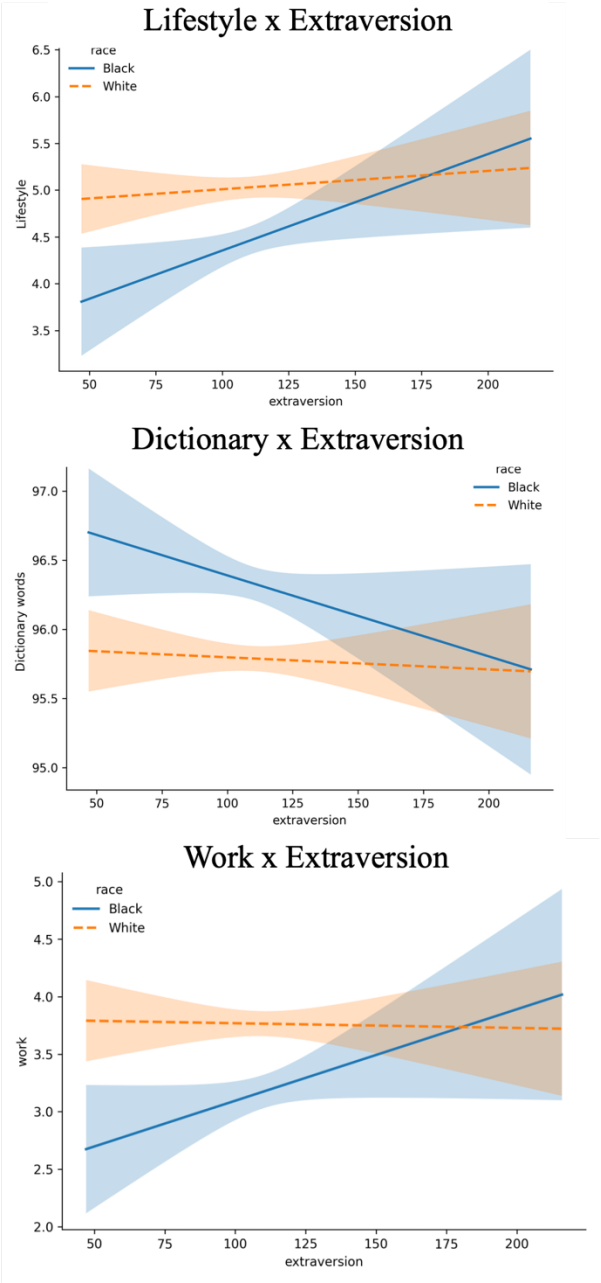
Model performance and bias across race for five FFM traits.



Note. Performance bias (ΔMAE) is interpreted alongside true-score variance differences; dot fill = per-group significance; bold line = significant racial difference between groups

Figure 2

Significantly different associations between extraversion and LIWC by race



Black/White Differences in Language-Based AI Modeling of Personality

Figure 3

Significantly different associations between FFM traits and LIWC by race

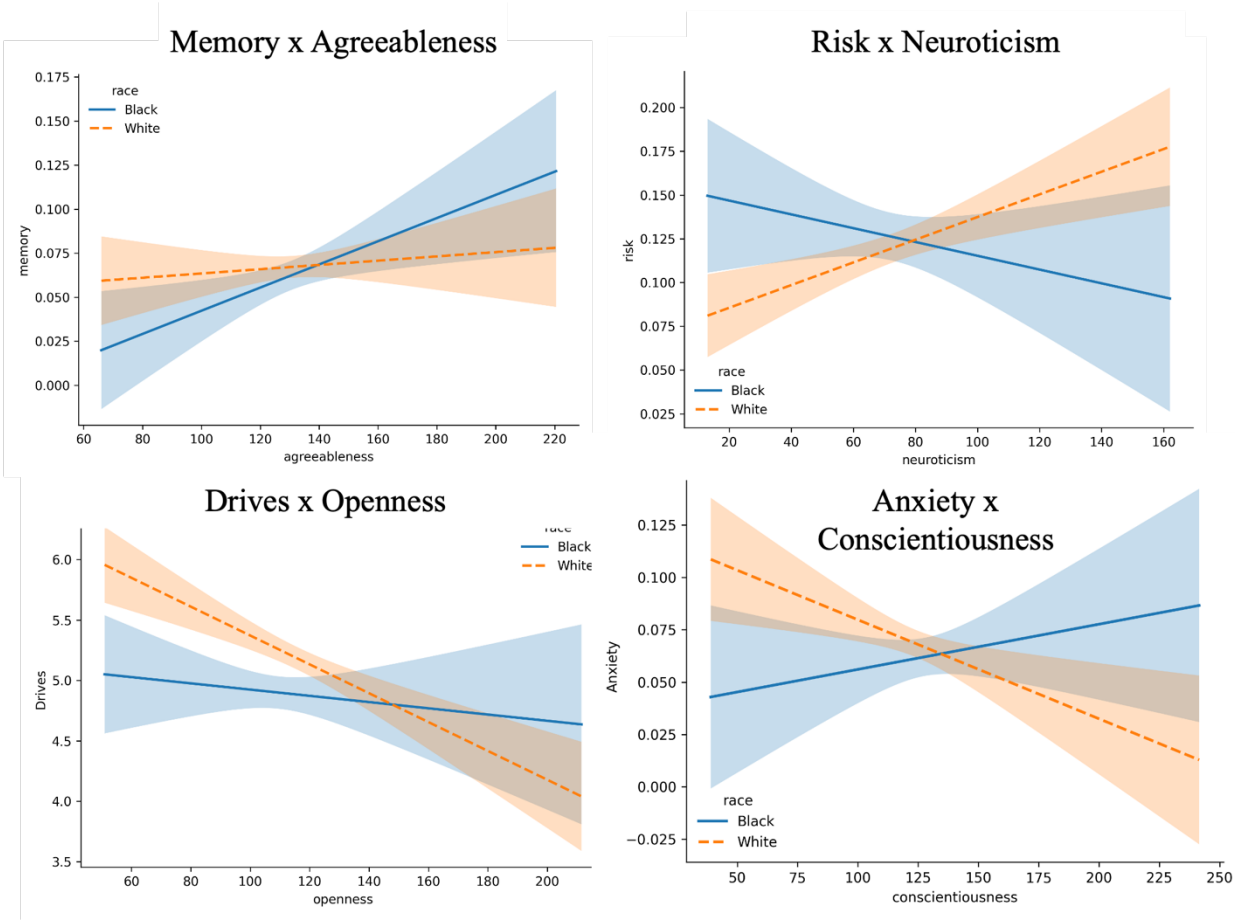


Figure 4

Significantly different associations between FFM traits and topics by race