Black and White Older Adults and Language-Based AI Modeling of Personality from Life Narrative Interviews

Tu Do, Tong Li, Mehak Gupta, and Joshua R. Oltmanns

Author's note:

Tu Do, M.A., 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.

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

The scientific study of five-factor model (FFM) personality traits began with the study of language. Now, language-based AI assessment of personality holds promise for the future. However, AI has a history of bias towards African-Americans, and modern language-based personality modeling should be examined for possible biases across Black and White research participants. To our knowledge, only one prior study has done so. The present study builds on Oltmanns et al.'s (2025) language-based personality models from $N = 1{,}405$ life narrative interviews with community older adults, this time examining the effects by race (Black and White American older adults). Personality was modeled using LIWC, BERTopic modeling, and fine-tuning of the RoBERTa language model (c.f., Oltmanns et al., 2025) for FFM personality traits. Results indicate significant differences in associations between personality and LIWC and BERTopic variables across race. Further, some fine-tuned RoBERTa language models for personality maintain relatively strong predictive performance for both groups, while others demonstrate a drop in predictive performance for Black participants (e.g., extraversion). Moreover, several models show mean-level prediction differences across groups, at small-to-moderate effect sizes, and small calibration errors indicating differences in predicted variance across groups. The findings provide important perspective regarding the language modeling of personality across race in American older adults and indicate there is work to be done to ensure fair performance of language-based AI modeling of personality across race. Implications for advances of language-based AI modeling of personality are discussed.

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

Black and White Older Adults and Language-Based AI Modeling of Personality from Life Narrative Interviews

The Five-Factor Model (FFM) of personality or the Big Five Model originated from the lexical hypothesis, which states that the most important personality traits are encoded in language (Allport & Odbert, 1936; Galton, 1884). Over the past one hundred years, a large body of emerging literature has supported the validity of five broad domains of personality (neuroticism, extraversion openness, agreeableness, and conscientiousness) that emerged from the initial lists developed by Allport found in the English dictionary (Goldberg, 1993; John, 2021). More recently, advances in artificial intelligence (AI) are making it possible to return to the lexical focus on personality assessment (J. R. Oltmanns et al., 2025; Park et al., 2015). However, almost no research to-date has examined Black-White racial differences in new language-based AI personality assessments.

The overwhelming majority of personality assessment research relies on self-report. This is problematic because of documented limitations of sole reliance on self-report (Paulhus & Vazire, 2007). For example, people may respond in socially desirable ways that are not true to their real personality traits. People may also have limited insight into certain areas of their personality (Vazire, 2010) and meta analyses demonstrate that self-other agreement on personality traits varies across types of traits and types of raters (Connelly & Ones, 2010), although some of this discrepancy may be related to trait variance (Mõttus et al., 2014). As a result, psychologists strive to incorporate multimethod assessment in research and clinical practice (APA Task Force on Psychological Assessment and Evaluation Guidelines, 2020). AI-based assessment methods may finally provide a scalable and more behavioral method of

integrating multimethod assessment into research and practice (Brickman et al., 2025; Kjell et al., 2023), overcoming the limitations of sole reliance on self-report.

However, very little research has considered the differential performance of AI based modeling of personality across race. And just as with self-report research, it is imperative that AI-based assessments are validated across diverse populations. AI already has a history of bias against Black Americans (O'Neil, 2016) and AI-based personality assessments should be validated on Black Americans before they are used with Black Americans in research and clinical practice. That is, researchers and clinicians should be aware of differential performance of test scores across populations of interest. Thus, the goal of the present study is to examine prior personality-based AI models developed in a representative community sample with relatively large representations of both White and Black older adults for racial differences in AI model performance and language-based differences related to personality.

Only one study to date has systematically evaluated the performance of language-based AI—in this case modeling only depression—across white and black research participants. In this case, 868 adults (n = 434 White, n = 434 Black) were recruited from social media and the language from their personality was modeled from their social media posts (Rai et al., 2024). It was found that across race, even the most widely cited linguistic feature of depression ("I" pronoun usage) did not transfer to Black participants. There were also several other topics in the language corpus as a whole that did not equally relate to depression for Black participants compared to White participants. For example, a topic related to describing feelings of worthlessness was only related to depression for White participants and not for Black participants. A machine learning model originally trained on Facebook status updates to predict depression scores using Linguistic Inquiry and Word Count software and BERT embeddings was

broken down by race. Models trained on only White participants worked best for White participants *and* for Black participants (Pearson *r* values ranging from .16 to *r* = .39). The models overall, whether trained on White or Black participants, worked less well for Black participants, with Pearson *r* values ranging from .06 to .13. These findings are important for continued research into language-based AI modeling of personality traits because they indicate findings at the total level—even some of the strongest linguistic features—may not be associated with depression (or personality traits) across equally across racial groups.

Oltmanns et al. (2025) examined personality prediction from language expressed in life narrative interviews (on average 20 minutes). Interviews were completed with *N* =1,409 community older adults. The RoBERTa language model was fine-tuned (i.e., retrained) on the language from the life narrative interviews and NEO-PI-R personality scores for the FFM domains. It was found that 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 a median 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 a median of .41 (extraversion and openness). These findings are remarkable for the size of the effects, which approach those that are considered large 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/African-American participants, testing of racial differences in that paper was outside of the scope of the research question.

Despite the importance of research showing the differences in language-based AI modeling of personality across race, there is a significant gap and lack of research in this area. It

is imperative that this issue is studied to examine model fairness and generalizability. To our knowledge, this is the first study to examine differences across Black and White adults in using fine-tuning of language AI models from life narrative interviews to predict personality traits. Research questions that are addressed include: Do language-based personality models perform differently across racial groups? Are models trained on one racial group less accurate when applied to another? What factors might explain performance differences? This study takes an essential broad step towards examination of fairness in language-based AI personality modeling.

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://github.com/AI-for-Health-Data/OCEANprediction). 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 dropped, resulting in 1,363 individuals. We performed a hybrid matching approach match the demographics of the Black and White samples as much as possible. Specifically, we implemented exact matching on education level and selected White participants with the most similar one-hot encoded annual income profile for each Black participant, yielding matched samples with a balance of key socioeconomic variables. This resulted in 450 matched Black-White pairs, with $M_{\text{age\_White}}$ = 60.1 SD = 2.9 (58.4% female). However, participant sample sizes in both groups vary across traits due to missing values.

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.

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., (2025). 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.

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 Black and White participants. 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 Black and White samples.

Fine-Tuning of RoBERTa

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

The RoBERTa-large model, 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 size of 16 and 8 for training and evaluation, respectively. Learning rate of 2e-5 was selected for the best results in Oltmanns et al. (2025).

Given RoBERTa's 512-token limit, a sliding window approach was used to input smaller chunks of the long, life narrative texts and calculate the mean of the predicted score from all chunks. To capture the full context of each life narrative, we averaged the classification embeddings of all chunks from the last layer of the RoBERTa to create a single embedding per participant, which was used to predict the final score with a feed-forward network.

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). 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, including tests of 1) accuracy, 2) performance differences, 3) directional bias and 4) calibration.

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.

Accuracy, Discrimination, and Distributional Comparisons. Model performance between Black and White samples was compared using several metrics:

Accuracy and Discrimination. 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 error magnitude (R^2) and ranking individuals (r).

Mean Bias. Mean prediction bias for each group was calculated as the mean of true score subtracted by predicted score. Positive results indicate underprediction while negative results indicate overprediction.

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.

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 true score minus predicted 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 underprediction for Black participants relative to Whites and $b_1 < 0$ indicates overprediction 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.

Results

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

LIWC

LIWC scores for the full sample are presented in Oltmanns et al., (2025). 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). The absolute value of *z* ranged from 2.58 to 2.94. These significantly different associations are displayed visually in Figure 1 (three from extraversion) and Figure 2 (one from neuroticism, agreeableness, conscientiousness, and openness each).

For the Black participants, there were 50 significant correlations of personality trait scores with LIWC scores. In the White participants, there were 59 significant correlations of personality trait scores with LIWC scores, with 16 of them being significant in both samples. All correlations are displayed in Supplemental Table 2. Effect sizes of the correlations between personality traits and LIWC variables were small according to subjective guidelines (Cohen, 1992): the median value was $r = .13$ for both groups.

BERTopic

Topics for the full sample are presented in Oltmanns et al. (2025). 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 $p < 0.01$ (Supplemental Table 3; Figure 3).

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]). None of the significant topic--personality correlations were similar across race.

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' median R^2 values across all training sets best predicted openness (e.g., White $R^2 = .20$, Black $R^2 = .13$). These sizes (with corresponding

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). Oltmanns et al. (2025) considered them large given the context and characteristics of the test (e.g., it is a true multimethod effect).

For White participants, RoBERTa predicted extraversion second best (median 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 (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 (median R^2 = .06 for both Black and White), followed by neuroticism (median R^2 = .05 for both Black and White). These two models worked equally well for both Black and White test groups, indicating no large significant difference in RoBERTa model performance by race across agreeableness and neuroticism. As in Oltmanns et al., (2025), the conscientiousness model showed the least predictive value, with a negative median R^2 value for the White test sets and .01 for the Black test sets. Median Pearson *r* values were .13 and .18 for White and Black, respectively. This indicates that the RoBERTa model worked least best 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 $bias_{White}$ = -0.09; pooled mean $bias_{Black}$ = 0.06).

Despite the racial differences in error magnitude, model fit was similar across groups. Correlations between predicted and true FFM scores were moderate (e.g., r = .28/.29) and statistically significant. 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, but neither was statistically significant, suggesting no over- or under-dispersion of predicted scores.

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 = -.16), nor in RMSE values. Mean prediction bias indicated small overprediction for the White sample and underprediction for the Black sample (pooled mean $bias_{White}$ = 0.04; pooled mean $bias_{Black}$ = -0.07).

Although model fit was relatively weak and correlations between predicted and true scores was low but significant for both groups, the model predicted significantly 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).

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 Blacks, 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 = .01), suggesting overprediction for Blacks. The magnitude of this effect was small (partial R^2 = .01), indicating significant but small directional bias.

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. 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 $bias_{White}$ = -.02; pooled mean $bias_{Black}$ = -.06). 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. 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 Blacks 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. This indicates underdispersion for White group openness predictions—Predicted scores varied less than true scores for Whites. 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 Blacks and overprediction for Whites, slightly larger bias for Black group (pooled mean $bias_{White}$ =.01; pooled mean $bias_{Black}$ = -.02). Model fit was the same for both groups (R^2 = .06). Correlations between predicted and true agreeableness scores were similar but significant (pooled r_{White} = .26 and pooled r_{Black} = .28). 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 = .01), indicating a small directional bias.

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.

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 both groups, slightly larger bias for White group (pooled mean $bias_{White}$ = -.15; pooled mean $bias_{Black}$ = -.13). Model fit was small poor in both groups, with negative R^2 values. This dovetails the findings of Oltmanns et al., (2025). Correlations between predicted and true conscientiousness scores were small but significant (pooled r_{White} = .14 and pooled r_{Black} = .19). 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 relatively large (partial R^2 = .06), indicating moderate 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. This indicates overdispersion for the White group conscientiousness predictions—Predicted scores varied more than true scores for Whites. Thus, results indicate statistically significant evidence of both calibration and directional bias in conscientiousness.

Discussion

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

Advances in AI have exciting implications for the improvement of psychological assessment. In particular, AI can assess behavioral features such as language in routine activities to increase the validity and scalability of validated assessment. However, little-to-no research has examined whether language models of personality operate equally across 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.

There are several ways to evaluate the comparative performance for Black versus White participants. One is pure predictive performance in the testing data separately for each race. In these tests, as judged by R^2, the degree of error from model prediction and ground truth, and Pearson *r*, rank order alignment of model prediction and ground truth, the model worked similarly for neuroticism, openness, agreeableness, and conscientiousness—with some small differences for example in the trait openness, R^2 = .20 for White versus .16 for Black and Pearson *r* = .46 for White versus .41 for Black. In contrast, we noted a more stark difference in predictive accuracy and rank order alignment for extraversion. While R^2 was .14 and Pearson *r* of .41 for extraversion in the White test set, the model predictions were only slightly better than chance and only aligned modestly in rank order in the Black test set (R^2 of .01 and Pearson *r* of .17).

Error-based metrics (RMSE and MAE) showed modest racial differences across neuroticism and openness, with Black sample showing slightly smaller absolute errors than White sample. These reduced errors coincided with significantly less variance in the true scores for the Black group, as evidenced by significant Levine tests. At first glance, this appears to support the validity of the performance of the Black model in the test data. However, MAE and RMSE are sensitive to the dispersion of the outcome variable; therefore, this may simply reflect

the lesser variance in the Black group, which would lead to less variable predictions. Indeed, across several traits, lower errors for Blacks co-occurred with smaller R^2 and *r,* suggesting that reduced error magnitude may reflect distributional properties rather than better predictive ability. The R^2 value provides better approximation of model prediction performance as opposed to distance from the score (as RMSE and MAE do).

When the Black group showed less error variance, as in openness and neuroticism, Black participants showed lower variance in true scores compared the White participants. Reduced variance constrains the range of prediction errors and can lower MAE and RMSE even when the model captures less variance in the outcome (i.e., lower R^2). For meaningful comparison of predictive accuracy across groups, the groups should have reasonably comparable variance in the criterion. This was still the case for the extraversion model (e.g., true score variance = 1.09 for White group and 0.89 for Black group, equaling a 81% ratio) and the agreeableness and conscientiousness models had just as much or more variance in the true Black group scores. However, the ratios were 71% and 74% for openness and neuroticism, respectively. Thus, it is possible but not obvious that the differing variance values may have caused the prediction problems for the Black group in openness and neuroticism.

In addition to predictive performance, it is also important to consider mean-level model predictions across groups. There was statistically significant (though small or moderate) directional mean level bias across race for all traits except neuroticism. Our interpretation is that there was small mean level bias in extraversion and agreeableness (e.g., R^2 was .01) but moderate mean-level bias in openness and conscientiousness (R^2 values were .09 and .06, respectively, with coefficients of b = .21 and -.13, respectively). These mean-level differences across race are notable and indicate that these models are not ready for use outside of research.

They indicate that, controlling for true mean level differences, the models underpredict openness and overpredict conscientiousness for African-Americans in our sample, to a moderate degree.

The RoBERTa model is a black box. However, we examined LIWC variables across race and conducted topic modeling to better understand why it may work better for certain traits for White Americans than for Black Americans. The most significant accuracy differences were in extraversion, which also showed the most significant differences in LIWC and BERTopic variables across race. Extraversion was significantly more positively related to discussion of work and lifestyle activities for Black participants but not White participants, and negatively related to LIWC dictionary (using more words) for Black participants but not White participants. Extraversion was also significantly more positively related to talking about teaching and graduate schoolwork for Black participants than for White participants. These LIWC and BERTopic variables converge in their common theme of work and lifestyle. This relatively large group of work and lifestyle related topics that were associated to extraversion for Black participants but not white participants may drive the significant accuracy differences that were seen for this trait (e.g., R^2 of .10 for White participants versus .01 for Black participants).

Other significant differences in trait associations with LIWC or BERTopic variables were not as pronounced but do provide some insight into linguistic differences across groups that may be important. For example, LIWC Risk (e.g., including words "secur*, protect*, pain, risk*) was associated with neuroticism for White but not Black participants and LIWC Memory (remember, forget, remind, forgot) and LIWC Anxiety (worry, fear, afraid, nervous) were more positively associated with agreeableness and conscientiousness, respectively, for Black participants to a greater extent than White participants.

Limitations and Future Directions

Although the present study is among, if not the first, examination of Black-White racial differences in language-based AI modeling of personality, it is not without limitations. The study has strong representation of Black Americans, but the full sample size had to be reduced because the original sample had a larger percentage of White Americans, which is natural given that White Americans are the majority population currently. However, it appears it will be critical for future research to continue pushing the sample size larger. Prior research has demonstrated that effect sizes may begin to stabilize at the level of a few thousand participants (Eichstaedt et al., 2020). It will be critical in future research to increase sample sizes to fully understand the potential of language-based AI modeling of personality. This may include online based video data collection, which could drastically increase the speed of data collection.

The present study is an important first step in the examination of the differential validity of language-based AI models for Black versus White participants. An important next step that could be completed with the current dataset is to examine other forms of construct validity. Fo example, convergent validity with informant reports by race and criterion validity for other important life variables by race (c.f., Oltmanns et al., 2025). These analyses will provide more insight into the meaningfulness of racial differences in language use associated with personality.

Generalizability

The results of this study apply only to Black and White American older adults. Future studies should examine their applicability to other age groups.

Conclusions

Although the advances in the possibilities of AI for psychological assessment are exciting, models need to be validated in the populations in which they will be used. This includes validating models across age, race, and gender, The present study is one of the first, to our

knowledge, to thoroughly examine the performance of language-based AI modeling of personality across Black and White American older adults. The results indicate that although there is relatively similar predictive performance of personality traits, there are differences in the utility of models across race that would indicate the models do not perform the same across race. Models of extraversion did not perform accurately for Black Americans as they did for White Americans and models of openness underpredicted mean levels for Black Americans compared to White Americans, controlling for true mean scores. These findings support how critical it is that AI models be validated across groups that they may be eventually used with in the real world. We stress the importance of ensuring a model has been validated with a specific group and is fully understood across groups before it is applied to that group. The findings here are initial evidence that researchers should examine racial group differences in their language-based AI models of personality to take initial steps for protection against unfair bias against minority groups moving forward.

References

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

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

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:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

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.

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

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.

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.

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2023). 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

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* [Preprint]. https://arxiv.org/abs/1907.11692

Mõttus, R., McCrae, R. R., Allik, J., & Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality*, *52*, 47–54. https://doi.org/10.1016/j.jrp.2014.07.005

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. (2025). Language-based AI modeling of personality traits and pathology from life narrative

Black/White Differences in Language-Based AI Modeling of Personality interviews. *Journal of Psychopathology and Clinical Science*. 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

O'Neil, C. (2016). *Weapons of math destruction*. Broadway Books.

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).

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/

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

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

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

Table 1. *Participant Demographics*

Variable	N_{White}	N_{Black}	$\%_{White}$	$\%_{Black}$
Age	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_{rue}	Var_{pred}	*F*	*df1*	*df2*	*p*
White	1.03	0.07	0.28	-0.09	0.1	0.01	1.14	0.13	8.61	1	889	0.003
Black	0.88	0.08	0.29	0.06	-0.1	-0.04	0.84	0.11				
Performance Bias	*MAE*	*ΔMAE*	*p*	*Cohen'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'd*								
White	0.77	-0.05	0.3	-0.16								
Black	0.72											
Directional Bias	b_1	*p*	R^2									
	-0.05	0.005	0.01									
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'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'd*								
White	0.74	0.02	0.74	0.02								
Black	0.76											
Directional Bias	b_1	*p*	R^2									
	0.08	0.002	0.01									
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			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	0.06									
Calibration	Slope	*p*										
White	0.53	0.01										
Black	0.84	0.43										

Figure 1

Significantly different associations between extraversion and LIWC by race

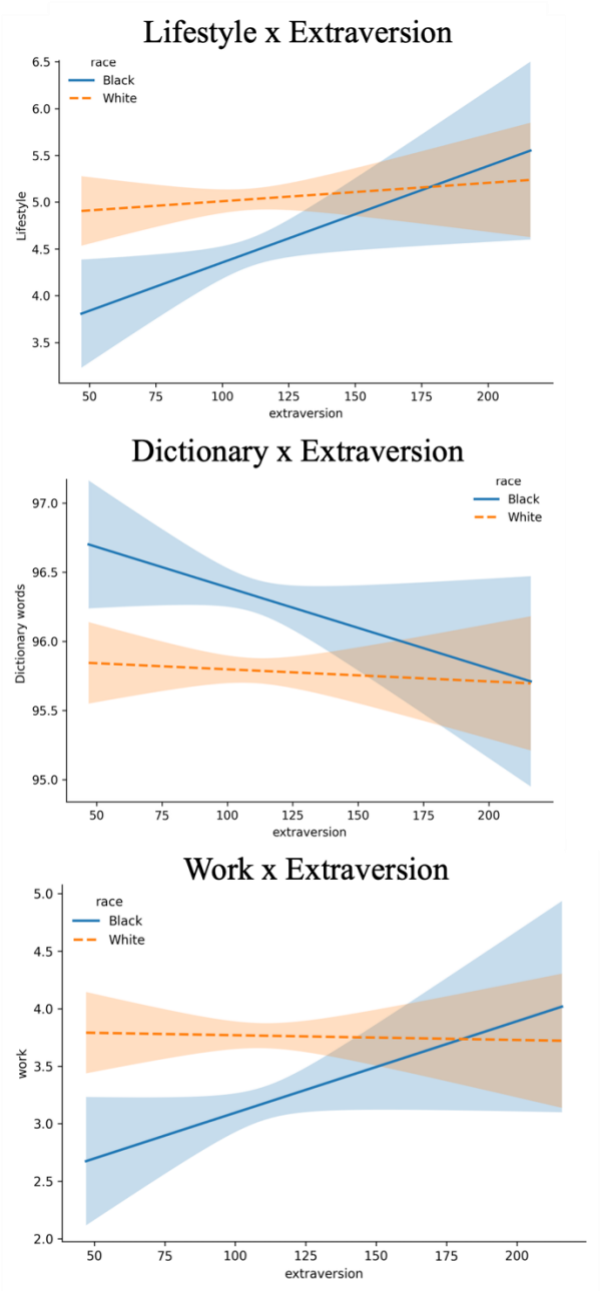
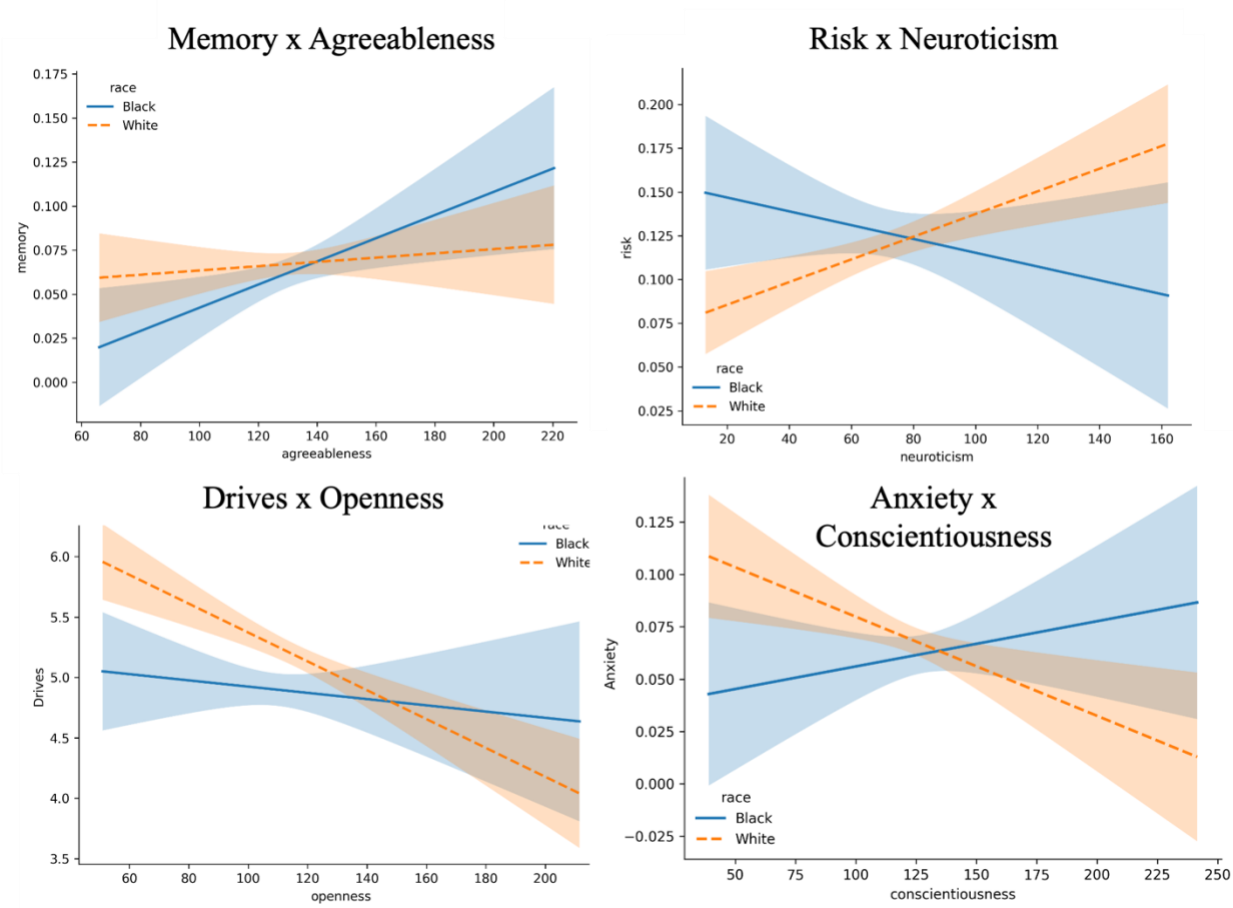


Figure 2

Significantly different associations between FFM traits and LIWC by race



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

Figure 3

Significantly different associations between FFM traits and topics by race