Language-Based AI Modeling of Five-Factor Model Traits and Personality Pathology from Life Narrative Interviews

Joshua R. Oltmanns[1], Ritik Khandelwal[2], Jerry Ma[2], Jocelyn Brickman[3],
Tu Do[1], Rasiq Hussain[2], and Mehak Gupta[2]

[1] Department of Psychological & Brain Sciences, Washington University in St. Louis,
[2]Department of Computer Science, SMU, [3]School of Psychology, Xavier University

Author's note:

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.

Analysis was supported by the high-performance computing system and a seed grant to the first and last authors provided by the SMU O'Donnell Data Science and Research Computing Institute.

The authors would like to thank the participants and researchers of the St. Louis Personality and Aging Network (SPAN). Their data collection efforts enabled this research and many other important investigations into personality, psychopathology, and health.

Correspondence should be addressed to Joshua R. Oltmanns, Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO. Email: j.oltmanns@wustl.edu.

Abstract

Personality disorders (PDs) are at a crossroads in classification and conceptualization. Advances in artificial intelligence (AI) and natural language processing hold promise for clarifying PD models and improving research methodology, understanding, and ultimately clinical treatment. This study uses language for modeling personality and personality pathology. A representative community sample of $N = 1{,}409$ older adults across the St. Louis region (33% Black, 65% white, 2% other) completed a life narrative interview from which language was used to train and test language models of personality based on scores from the NEO-Personality Inventory-Revised (NEO-PI-R) and the Structured Interview for DSM-IV Personality (SIDP-IV). Criteria measures were used for multi-method construct validation of the language models including self-report measures of physical functioning and depressive symptoms and informant-report measures of personality, general health status, and social functioning. Language from life narrative interviews was modeled to identify personality through fine-tuning the parameters of the RoBERTa language model, BERTopic topic modeling, and Linguistic Inquiry and Word Count. Fine-tuned RoBERTa models predicted personality scores in testing data above $r = .40$. Language in life narrative interviews supports the semantic similarity of the five-factor model (FFM) personality trait domains more than DSM personality disorder categories, for which only borderline pathology had support. The language-based FFM scores were supported by multi-method criteria correlations including informant-report personality scores in the testing data. Findings support dimensional conceptualization of personality and demonstrate the promise of language-based AI to refine conceptual frameworks of PD and provide automatic personality assessment and prediction in research and clinical practice.

Keywords: personality disorders, five-factor model, personality, AI, NLP, LLMs

Language-Based AI Modeling of Five-Factor Model Traits and Personality Pathology from Life Narrative Interviews

The field of personality disorders (PDs) is at a crossroads. New dimensional models are poised to take over traditional categorical PD classification models and have already done so in the *International Classification of Diseases—11th Revision* (ICD-11) (World Health Organization, 2025). Unfortunately, differences in perspective over these dimensional models has led to much debate that can cause a standstill. New methods of assessment leading to useful tools for future research and practice may contribute to forward progress. In particular, rapid innovations in AI allow new avenues for assessing PD through language. These methods have shown promise and may be primed to improve assessment and knowledge of PD.

To date, the field has relied almost exclusively on diagnostic interviews and self-report questionnaires for assessment of PD. While there are plenty of valid and reliable self-report assessments, sole reliance on one method is precarious and not aligned APA assessment guidelines (APA Task Force on Psychological Assessment and Evaluation Guidelines, 2020). Self-report has limitations and the PD field should strive for new multi-method assessment tools that can increase the validity of assessment. Language can help bypass self-report limitations, adding richness to assessment through recognizing nuanced language features associated with PD. Further, language-based assessment simultaneously provides the opportunity for automatic implementation of assessment into routine procedure, which could address issues such as the inadequate frequency of formal assessment by practicing clinicians (Hatfield & Ogles, 2004).

Language-Based Assessment

Psychologists have long been interested in language (Sanford, 1942). Modern dimensional personality framework stems from the study of language (Allport & Odbert, 1936).

All personality-relevant words from the dictionary were factor-analyzed repeatedly over the years to identify five broad factors (Digman, 1990). Later in the 20th century, programs and rating systems were developed from language to simulate therapists and assess psychological states (Gottschalk & Gleser, 1969; Stone et al., 1966; Weizenbaum, 1961). The Linguistic Inquiry Word Count (LIWC) software (Pennebaker et al., 2003) was developed in the 1990s. LIWC uses a “bag of words” approach that counts word frequencies in a document. It then uses a “top-down” approach to score language within broader psychological processes (e.g., positive and negative emotions, cognitive processes) and other linguistic categories (articles, singular, and plural pronouns) (Boyd et al., 2022).

LIWC has been a groundbreaking tool for improving understanding of personality. Research using LIWC across various language samples including daily diaries, class assignments, and academic article abstracts, and contexts including in daily life, has shown that neuroticism is associated with negative emotion words, fewer positive emotions words, and first person singular usage; extraversion is associated with positive emotions words; social-process words, word count, lower complexity; and agreeableness is associated with positive emotion words and fewer negative emotion words (Mehl et al., 2006; Pennebaker & King, 1999; Tackman et al., 2019). Borderline PD has been associated with negative emotion words and paranoid PD with angry words (Calabrese et al., 2024; Entwistle et al., 2023), and grandiose narcissism with swear words, second-person pronouns, and negatively with anxiety/fear words (Holtzman et al., 2019). Although statistically significant, these associations are usually small (e.g., with a range of $r = .10$ to $r = .16$).

LIWC does have limitations (Jackson et al., 2022; Lawson & Matz, 2022). First, scores are based on human judgment. The top-down approach relying on human-coding of psychological constructs and dividing words into pre-existing categories may inadequately represent or miss important words. It is also difficult for human raters to code all possible uses and meaning of all words. Second, LIWC pre-existing "dictionaries" used to assess linguistic features may not include certain words and word frequency may not adequately capture the importance of a word. Third, LIWC does not understand word context and is unable to pick up on certain figures of speech or grammatical functions such as irony or negations. For example, "mad" might describe anger, mental illness, or quantity (slang). These limitations are now addressed by advanced natural language processing (NLP) techniques.

Machine Learning and Natural Language Processing

In the 2010s, machine-learning-based methods for improving NLP were refined. Word2vec introduced an efficient version of "word embeddings," which use machine learning to convert words in text documents into vectors (i.e., strings of numbers) (Mikolov et al., 2013). These embeddings provide a quantitative, multidimensional representation of a word's meaning in relation to other words in the text. Techniques from this time usually produce a fixed 100-to-300-dimensional vector per word. However, the machine learning-based word embedding framework ELMo (Embeddings from Language Models; Peters et al., 2018) introduced *contextualized* embeddings, where a word has a different embedding depending on the other words in its proximity. In this way, words with multiple meanings (e.g., "mad") would have different embeddings based on their usage.

The transformer deep learning model architecture improved the quality and speed of NLP such that it enabled large language models (LLMs) as we know them today (Vaswani et al., 2017). Embeddings had previously been used in deep learning frameworks one word at a time, sequentially, to predict the next word. Transformers implement "self-attention" and allow attention to all words in an input text *simultaneously*, understand the relationships between all words, and even speed up the modeling process (Brickman et al., 2025). LLMs use the transformer architecture to pretrain embeddings on vast amounts of data. For example, one of the early influential models is Bidirectional Encoders for Representations of Transformers (BERT), which was trained using word masking and next sentence prediction tasks on the entirety of English Wikipedia and BooksCorpus (Devlin et al., 2019). This gives LLMs a stronger quantitative understanding of language and makes them more reliable than embeddings trained on smaller amounts of data. But transformer models are also contextualized and dynamic–they produce embeddings during the modeling process based on the text they are provided—capturing the meaning of each word in relation to all other words in the text.

Studies of personality using transformer models or contextualized embeddings from transformer models have been infrequent (Jain et al., 2022; Mehta et al., 2020). One study pulled 9,000 sentences from social media that were deemed PD-related using dictionaries built from prior research findings, for example negative emotion words, swear words, and first person pronouns (Jain et al., 2024). BERT and variations of BERT (RoBERTa, DistilBERT) were fine-tuned to predict a PD label defined by usage of two words in the text that were deemed related to PD through a PD-related language corpus and ratings from two psychologists. Across cluster B disorders, BERT had the best accuracy (.750). However, the sample size was not listed, the dictionaries for the disorders overlapped significantly (*r* ~.80-.90), and several disorders had very low accuracies (e.g., histrionic F1 = .396). Finally, the identification of PD was not stringently validated, which has significant limitations in terms of diagnostic reliability. However, problematic PD labeling in datasets, such as using self-disclosure as a dependent variable to train models, is often relied on out of necessity in social media language-based studies of personality and mental disorders because there are no other validated measures completed by the participants in the sample. Other limitations of prior studies include predicting categories of outdated personality models or categorizing dimensional models.

Some transformer-based modeling studies have predicted personality from social media generating continuous outputs, which is more in line with current PD assessment. Lynn and colleagues (2020) predicted FFM personality domains from Facebook posts and found stronger predictive power (average $r = .56$). Personality has also been predicted with transformer models using text from Reddit with 1,105 participants (no demographic information collected) who completed an FFM personality questionnaire (Simchon et al., 2023). Fine-tuned BERT showed prediction of the FFM domains ranging from $r = .26$ (extraversion) to $r = .39$ (openness), with a median of $r = .35$ (Simchon et al., 2023). These studies provide a strong foundation for transformer-based LLMs in personality assessment. However, despite some evidence that models built on social media language may replicate to spoken language (J. R. Oltmanns et al., 2021), reliance on social media language is limited in terms of potential clinical applications, diagnostic validity, and relevance to spoken language, which is the preferred method of gathering information in clinical settings.

In addition to use of transformer-based models for identifying PD, topic modeling is useful for understanding personality from language. Topic associations with FFM personality traits have been examined previously from social media data (Park et al., 2015). There were strong face valid representations, for example some topics most related to introversion included computers and reading, and some topics most related to extraversion included partying and love. More recently BERTopic was introduced, which combines topic modeling with transformer-based modeling (Grootendorst, 2022). This enables topic modeling using contextualized embeddings. BERTopics can be correlated with constructs of interest. To our knowledge, BERTopic has not yet been used to examine PD. Examining topics in combination with fine-tuned LLM embeddings and LIWC scores may provide insight into personality.

The Present Study

To our knowledge, no prior study has fine-tuned a transformer model on spoken language for personality and personality disorders. The present study makes advances in several other ways: 1) most studies using LLMs for personality assessment have relied on social media language samples, which differ substantially from clinically spoken language; 2) most prior studies have used categorical classification of personality or PD, which are more limited than continuous scoring and inconsistent with current personality assessment techniques in psychology; 3) all prior studies have focused on only one conceptualization of personality or PD, thereby providing no opportunity to learn about differences between traits and disorder from language; and 4) most prior studies do not report demographics or use predominantly white samples.

The present study addresses these issues. First, it is imperative to train, test, and evaluate relative validity of models from spoken, in-person interview language. The present study will develop models from spoken language used in life narrative interviews that were completed in person, face-to-face, in a manner that is largely consistent with an initial clinical assessment. Second, the present study will develop language models that are based on dimensional scoring and provide output that aligns with the dimensional conceptualization of personality. Third, the present study will train models on multiple conceptualizations of personality/PD (i.e., the FFM and the categorical DSM PD model), providing an opportunity to compare and contrast multiple PD models through language. Finally, the development of AI-related technology already has a history of bias against underrepresented communities and it is imperative that development of AI technologies moving forward be inclusive to make a best attempt towards recognizing and addressing bias as much as possible. The present study uses data collected from a representative community sample of *N* = 1,405 older adults in St. Louis, Missouri, USA that matches census data of the area. Although the present study does not investigate differences across race, representation of Black Americans (~33%) in addition to White Americans will lead to a language model that will be more representative of the community as a whole.

Method

The present study was not preregistered and the analyses were exploratory. We hope the knowledge gained from this study facilitates preregistration of future studies. These findings should be replicated and evaluated because of their exploratory nature. The code is available online (https://github.com/AI-for-Health-Data/OCEANprediction). All data are available on the Open Science Framework (https://osf.io/6pq7w/?view_only=651a190683d94e8e90ed4cd78ef7ce2f), with the exception of the life narratives which are not openly available for privacy reasons. However, the life narrative transcripts are available upon formal request and data sharing agreement. We report how we determined sample size, all data exclusions, all manipulations, and all measures in the study.

Procedure

Data come from the St. Louis Personality and Aging Network (SPAN), for which a representative community sample of 1,630 older adults was recruited across 100 square miles in the St. Louis area from 2007 to 2011. Listed phone numbers were used to contact households and the Kish (1949) method was used to identify targets for participation within households. Participants came to the laboratory and completed the life narrative interview, the Semistructured

Interview for DSM-IV Personality (SIDP-IV), the NEO-Personality Inventory-Revised (NEO-PI-R), and a battery of other measures related to personality and health (T. F. Oltmanns et al., 2014). The study was approved by the local 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). Participants were representative of the St. Louis area in terms of race and ethnicity and came from a broad range of socioeconomic status, with a slightly higher median household income compared to the 2008 median in St. Louis (the time the data were collected). Black men were oversampled for participation after initially lower participation rates compared to Black women and White men and women (Spence & Oltmanns, 2011).

Participants each nominated an informant who “knew them best” to also complete questionnaires about them and 90% of the 1,409 target participants had an informant (*N* = 1,264 informants). Informants were 68.1% female and 31.8% male, romantic partners (47.8%), family members (28%), friends (21.9%), and the remainder were neighbors, co-workers, or other. Informants were 66.9% White and 30.5% Black, and 2% other. Informants reported they had known the target participants for 32.5 years, on average (*SD* = 15.0). Informants mostly reported they knew the target better than anyone else (49.4%) or very well (42.4%), liked the target more than anyone else (49.4%) or very much (47.2%), and were closer than anyone else (51.0%), very close (41.9%), or somewhat close (6.0%).

Measures

Life Narrative Interview. The life narrative interview was adapted from McAdams (1993). Participants were asked to provide their life stories, beginning at age 18, and divide them into 3-4 chapters. At the conclusion, they were asked to list their best and worst characters, high and low points, and a turning point. Life narratives lasted 20 minutes on average.

NEO-Personality Inventory-Revised (NEO-PI-R). The NEO-PI-R (Costa & McCrae, 1992) is a 240-item measure of the FFM of personality. Items are rated on a Likert-type scale from *strongly disagree* to *strongly agree*. Scaled scores on the NEO-PI-R were created for each domain (extraversion, agreeableness, conscientiousness, neuroticism, and openness) for both targets and informants. If scales were missing 1 or 2 items, completed items were averaged. If they were missing more than 2 items, they were removed from the dataset. The NEO-PI-R has demonstrated strong psychometric characteristics in the SPAN study dataset including internal consistency, test-retest reliability, and criterion validity across both target participants and informants (J. R. Oltmanns et al., 2020; Wright et al., 2022).

Structured Interview for DSM-IV Personality (SIDP). The SIDP (Pfohl et al., 1997) is a structured interview assessment of the DSM personality disorder categories. The criteria for the DSM-IV PDs were rated *not present* (0) to *strongly present* (3). Trained interviewers were Ph.D., master's level, and undergraduate psychology students. Case conferences with the whole team including the PI on the NIH grants were used to maintain rater cohesion throughout data collection. For the present study, schizoid, borderline, and obsessive-compulsive PDs were used to represent clusters A, B, and C, respectively, because they represent theoretically distinct forms of pathology and each predicted important outcomes in SPAN (e.g., Powers et al., 2013). Inter-rater reliability of $N = 265$ re-rated interviews was $ICC = .67$ for a continuous total of the SIDP criteria (Gleason et al., 2012). The *ICC* for borderline PD was .77, for schizoid PD was .75, and for obsessive-compulsive PD was .62. Scores were created by summing the number of criteria for each PD rated as *present* (2) or *strongly present* (3).

Criteria measures. The RAND-36 Health Status Inventory (HSI) (Hays et al., 1998) was used to assess subjective physical functioning. The scale consists of 10 items assessing limitations of physical functioning especially relevant to older adults (e.g., “Does your health now limit you in these activities? If so, how much?”), with an example item being “Lifting or carrying groceries.” Higher scores indicate better physical functioning. The Beck Depression Inventory (BDI-II) (Beck et al., 1996) was used to assess depressive symptoms. The scale includes 21 items assessing symptoms related to depression over the past 2 weeks. Participants indicated the severity of their symptoms on each item on a scale from 0 to 3, with 3 being the worst. Higher scores indicate more depressive symptoms. Coefficient alphas for the HSI Physical Functioning scale and the BDI-II scale in the SPAN study are about .90 (Cruitt & Oltmanns, 2019).

Informants completed two measures of the target participants’ current functioning: The informant HSI (IHSI) was completed by informants about target participants’ physical and emotional health (Cruitt & Oltmanns, 2018). It included 10 items that were deemed easier to rate by an informant. An example item is “During the past 4 weeks, to what extent has his/her physical health [or emotional problems] interfered with his/her normal social activities with family, neighbors, or groups?” Items were rated on 5 or 6-point scales, for example from *poor* to *excellent*. Higher scores indicated better health. Coefficient alpha was .87. Informants also completed an eight item informant-version of the Social Adjustment Scale (Weissman, 1999). Example items include “How well has he/she been able to do his/her work in the last 2 weeks?” and “Has she/he had any open arguments with friends or relatives in the last 2 weeks?” Items were rated on a 1-5 scale with statements corresponding to the item content. Higher scores reflect poorer adjustment. Coefficient alphas was .68.

Analysis

Transcript Deidentification. Life narrative interviews were transcribed manually by Speechpad. Transcripts were deidentified with named entity recognition using the "en_core_web_trf" model from the spaCy python package. This model identifies words from 18 built-in categories. Entities in the transcripts recognized from the PERSON (people), FAC (budlings, airports, highways, bridges), ORG (companies, agencies, institutions), GPE (countries, cities, states), and LOC (non-GPE locations) categories were extracted and replaced with general terms (e.g. "person," "geopolitical location," or "organization"). Some terms deemed relevant to older adult personality, but not important for deidentification, within these categories were retained, for example "Vietnam," "JFK," "9/11," and "USA." Street addresses were removed and specific dates were reformatted to only include month and year.

Transcript Preprocessing. A preprocessing pipeline was tailored to 1,127 training files and 282 testing interview transcript files. Preprocessing tasks included the removal of text annotations related to the audio transcript (e.g., "crosstalk," "inaudible," or "silence") and conversion of all text to lowercase. Language associated with the interviewer was removed. Most interview recordings began with a statement of consent, which was removed. Stop words (e.g., "the," "is," "and") were retained. The processed data were subsequently saved into two new CSV files, one for the training data and one for the testing data.

Linguistic Inquiry and Word Count (LIWC). LIWC (Boyd et al., 2022) was used to score the texts for psychological processes and parts of speech using a top-down approach.

Deidentified texts were entered and 117 variables were computed. The emoji frequency variable was removed. Descriptives for these variables are provided in the Supplemental Materials.

Fine-Tuning of RoBERTa. We fined tuned the RoBERTa-large model (Liu et al., 2019) using the simpletransformers library (Rajapakse, 2024) for the regression task of predicting FFM personality traits and SIDP PD scores. RoBERTa-large is a model by Facebook AI trained on 1024 V100 GPUs for 500,000 steps. Using a batch size of 8,000 and a sequence length of 512, it was trained using the masked-language modeling objective on large, diverse text corpora. RoBERTa is an optimized version of Google's BERT language model (Devlin et al., 2019). While BERT was trained on English Wikipedia and the BooksCorpus dataset, RoBERTa was trained on 10 times more data, including the Common Crawl corpus. It was also trained using a dynamic masked language modeling approach as compared to a static one used for BERT. These additional training techniques led to a better language model.

The model was fine-tuned on the life narrative interviews with the following hyperparameters: a maximum sequence length of 512 tokens (max sequence length of the pre-trained model), a batch size of 16 for training, and 8 for evaluation. Learning rates of 2e-2, 2e-3, 2e-4, and 2e-5 were tested for optimization, and 2e-5 was selected for the best results across models. Because RoBERTa has a 512 token limit, but the life narrative texts were longer, a sliding window approach was used to input smaller chunks of the text.

RoBERTa predicts final score by taking mean of the predicted score from all the small chunks (windows). This strategy did not capture the full context of the life narrative because each small chunk is predicting score in complete isolation from other windows. To capture the full context of each life narrative, we used the classification embeddings of all chunks from the last layer of the RoBERTa. We then averaged these embeddings to create a single embedding per participant, which was used to predict the final score with a feed-forward network. Training time for one personality score in RoBERTa took around 2 hours and 15 minutes for the feed-forward network on NVIDIA A100 GPU computers. Details of the RoBERTa method can be found in Brickman et al. (2025). Code for both RoBERTa and the feed-forward networks is available on GitHub (https://github.com/AI-for-Health-Data/OCEANprediction).

Five-fold cross validation was used with three data splits – train, validation and a hold-out test set. The validation set was 5% of the train set selected randomly. After training the model on the train set, every epoch was tested on the validation set. Results from the validation set were used to decide if early stopping was necessary (i.e., if case performance did not improve for five continuous epochs). After training the model was saved with trained weights. The model was then put in evaluate mode and evaluated on the test set. We report all results based on the test set. The evaluation metrics included mean squared error (MSE), mean absolute error (MAE), and R-squared (R^2) score. Converted pearson *r* values and MSEs are reported in the Results. Full evaluation metrics with standard deviations can be found in the Supplemental Materials.

BERTopic. BERTopic (Grootendorst, 2022) was used to identify topics in the life narratives. BERTopic uses pretrained BERT embeddings to identify meaning. It uses the following process: Embedding sentences in documents, reducing dimensionality, clustering the embeddings into topics, and tokenizing and weighting the topics. In BERTopic, the default embeddings are at the sentence level from Sentence BERT (sBERT). Uniform Manifold Approximation and Projection (UMAP) was used to reduce the dimensionality of the embeddings, with the following settings: 15 nearest neighbors, 5 components, 0.0 min_dist, cosine metric. Out of multiple component values tested (2, 5, 10), five offered the best variance balance for meaningful clusters to the original embeddings while topic coherence and semantic structure were still reasonably facilitated. Ten components introduced noise and topic coherence decreased. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was used to cluster the resulting components. We tested a number of values (5, 10, 15, 30) of minimum cluster size and found that a value of 15 and a euclidean metric provided the best trade-off with lower Davies-Bouldin Index of 8.0 and higher Silhouette Score of 0.22—allowing smaller, fine-grained topics to emerge while minimizing fragmented clusters and not providing an excessive number of topics. The combination of cosine distance for UMAP and euclidean distance for HDBSCAN worked best in producing interpretable clusters rather than using cosine distance for both operations. All other settings were kept at their default values. Finally, class-based Term Frequency-Inverse Document Frequency (cTF-IDF) was used to tokenize the topics.

By default, BERTopic identifies one topic per document. In life narratives, it makes more sense to identify several topics within each life narrative (people are not discussing one topic in their life narrative). Therefore, we divided the transcripts into paragraphs, or utterances. Most participants had many utterances and BERTopic assigned one topic per utterance, allowing each participant to be associated with multiple topics. The probability was calculated of each participant discussing each topic. The maximum probability for each topic across all utterances for each participant was used as a single probability value for each topic per participant, and was then used to correlate with personality scores. Use of the average probability per topic per participant was also tested, and there were not major differences between approaches. However, our use of correlation between personality and topics assumes linear relations.

We used the topic probability vectors from BERTopic as features for predicting personality scores. For each participant, we created a 150-dimensional vector by aggregating the maximum probability over all utterances in their life narrative. This approach emphasizes topics discussed with higher intensity, assigning lower probabilities to less frequently mentioned or absent topics. We also experimented with a frequency-based approach, counting probability scores of one (i.e., definitely present) per participant across different sentences. We used this to calculate correlation of topics with every personality trait and three disorders. Our results showed majority overlap between topics found with the frequency and max probability methods. With the maximum probabilities, we trained a two layer feed-forward neural network (with layer dimensions of 32 and 64 dimensions, dropout rate of 0.2 to prevent overfitting, LeakyReLU activation (Xu et al., 2015) and the Adam optimizer with a learning rate of 0.01) using these vectors to predict personality scores.

Combined Model. The predicted scores obtained from fine-tuned RoBERTa, BERTopic embeddings, and LIWC scores were combined in a linear regression model to obtain final predicted personality scores.

Results

Descriptives for the NEO-PI-R and SIDP-IV scores are presented in Table 2. The NEO-PI-R FFM personality scores were continuous and normally distributed. The SIDP PD scores were significantly positively skewed and less variable than the NEO-PI-R scores. There was personality pathology present in the sample, including people who met full criteria for PD: 39 people met full criteria for OCPD, 5 for BPD, and 10 for SZPD. Conceptualizing the PDs continuously, 271 people had one or more symptoms of SZPD, 202 people had one or more symptoms of BPD, and 589 people had one or more symptoms of OCPD.

LIWC

LIWC indicated that participants used 2,359 words on average (*SD* = 1,857 words), and this distribution was positively skewed (skewness = 2.3, *SE* = .07). Participants' usage rates of "I" and "we" pronouns were similar to those presented in the LIWC descriptive statistics manual for conversations. Usage rates are presented as frequencies of total text: Participants used positive and negative emotion words (*M* = 0.6%, *SD* = 0.4%, range 0% to 2.8% and *M* = 0.4%, *SD* = 0.3%, range 0% to 2.0%, respectively). Swear words were used in 0.02% of the texts, on average (*SD* = 0.1%), range 0% to 1.3%. Participants' words related to social behavior (*M* = 2.4%, *SD* = 0.8%, range 0.4% to 8.0%), work (*M* = 3.5%, *SD* = 1.6%, range 0% to 11.6%), and health (*M* = 0.6%, *SD* = 0.5%, range 0% to 4.0%). A full list of descriptives for the LIWC variables is provided in the Supplemental Materials Table S1.

The top correlated LIWC features with the personality scores are presented in the Supplemental Materials Tables S2-S7. Correlations were small (i.e., in the *r* = .10 to *r* = .20 range) but often face valid: Higher neuroticism scores were associated with higher negative emotion words, higher extraversion scores were associated with positive emotion words, openness was most associated with curiosity words, agreeableness was associated with prosocial and affiliation language, and conscientiousness was most associated with work language. The most strongly correlated features with borderline overlapped with neuroticism (physical language, negative emotions and tone, personal pronouns). SZ was negatively correlated with words that were positively correlated with extraversion and agreeableness.

Associations between the *patterns* of correlations of personality and LIWC scores are presented in Supplemental Table S8. Extraversion's LIWC feature pattern was strongly negatively correlated with SZ's LIWC feature pattern. Neuroticism's was strongly positively correlated with BD's and negatively correlated with conscientiousness. Openness's LIWC feature pattern was most correlated with SZ (negatively), agreeableness's pattern was most strongly correlated with OC (negatively), and SZ and BD LIWC feature patterns correlated highly, while OC's LIWC feature pattern did not correlate strongly with SZ's or BD's.

BERTopic

BERTopic modeling identified 149 topics (Supplemental Materials Table S9). Figure 1 shows the top 10 most prevalent topics in the life narrative interviews. Participants most frequently discussed family, life and death, impactful people, marriage years, Christianity, military and war, and drinking and drugs. The full table of correlations between topics and personality scores is presented in Supplemental Table S10. Topics with the strongest correlations with the personality scores are presented in Figure 2. All were within the absolute value effect sizes of $r = .08$ and $r = .14$ and significant at $p < .001$. BD was significantly correlated with seven topics, neuroticism and agreeableness were significantly correlated with six topics, OC with five topics, openness with four topics, extraversion and conscientiousness with three topics, and SZ with two topics.

Many topic correlations were also face valid: Borderline and neuroticism had correlations with several negative health outcomes including alcohol and drugs, physical health problems, and medications. Agreeableness was negatively associated with the military and schizoid was negatively associated with work. Openness had positive correlations with talking about work and arts/music There were also interesting unforeseen correlations: BD was associated with discussion of inner peace and SZ was relatively strongly associated with the military/war topic.

Predictive Models

Table 3 shows the effect sizes of the predictive models of personality. The blank values indicate the model did not provide positive values for predicting the respective personality score.

LIWC features were associated with personality at a small effect size: The median value was *r* = .26 for the FFM scores and *r* = .14 for the PD scores. This aligns with the LIWC correlations presented in the supplemental materials. BERTopic probabilities were also small effect-size predictors of personality: The median value for the BERTopic features was *r* = .17 for the FFM scores and .14 for the PD scores.

Fine-tuned RoBERTa embeddings were the relatively strongest predictors of personality: *r* ranged from .32 (agreeableness and conscientiousness) to .45 (openness). The median pearson *r* value for the RoBERTa models was *r* = .39 for the FFM scores. However, the fine-tuned RoBERTa model was only able to positively predict BD could not positively predict SZ or OC. Training as a multi-label classification model with 6-7 labels did not help obtain positive predictive values. Training the PD models as classification tasks did not help obtain positive results.

RoBERTa embeddings for the CLS tokens (summary embeddings) are visualized in 2D space in Figure 3. These plots indicate associations between general semantic representations of life narrative interview language and FFM personality traits—for example, people who were higher on neuroticism provided more semantically similar life narratives than people who were lower on neuroticism. The openness plot indicates nonlinearity in the relations between openness and semantic meaning of the life narratives—the "U" shape with the top ends bending closer to one another than to the average scorers in the bottom of the U indicates that people higher and lower on openness had more semantic similarity in their narratives than average scorers.

The CLS embeddings for the PD constructs were less strong. No relation is apparent between the semantic representations of the life narratives and levels of SZ and OC. For BD, there is a modestly detectable relationship between semantic meaning and BD—this can be identified by the smaller cluster of darker CLS embeddings on the right of the BD plot. These findings for PD align with the fact that SZ and OC models were not positively predictive and although the BD model was positively predictive, it was less predictive than the FFM models.

The results indicate that the predictive validity of the combined language model was largely unchanged when LIWC, BERTopics, and RoBERTa were combined, but decreased for conscientiousness and increased for BD. The median *r* value for the combined models was .41 for the FFM scores and .14 for the PD scores (of course, the SZ and OC combined models only included LIWC and BERTopic predictors).

To examine cross-sectional multi-method convergent, discriminant, and criterion validity, the predicted language model personality scores for one testing fold from the fine-tuned RoBERTa models were correlated with informant-reported FFM personality domains and borderline pathology, self-reported physical functioning and depressive symptoms, and informant-reported general health status and social adjustment (Table 4). The fine-tuned RoBERTa models showed convergent and discriminant validity with informant-reported FFM domains. The fine-tuned RoBERTa BD model did not show convergent validity with informant-reported BD pathology. However, it did show some convergent validity in that it showed the FFM BD trait profile: it correlated positively with informant-reported neuroticism and negatively with informant-reported agreeableness, and conscientiousness. The fine-tuned RoBERTa models showed significant correlations with self-reported life criteria at small-to-moderate effect sizes, except for agreeableness. These results support the construct validity of the fine-tuned RoBERTa language models of FFM personality and modestly support the fine-tuned RoBERTa language model of BD.

Discussion

Results from the present study show promise that language may eventually be used as a valid personality assessment method (c.f., Pennebaker & King, 1999). Fine-tuned RoBERTa models were the strongest predictors of personality scores, ranging from *r* = .26 (BD) to *r* = .45 (openness), with a median of *r* = .36. LIWC and BERTopic provided smaller associations with the personality scores. Combined models most often did not show improved validity, with the exception of the BD model. This indicates that fine-tuning language models may prove more useful as predictive tools for personality assessment than LIWC and topic modeling. However, LIWC and BERTopic provide essential substantive information about language use and personality.

Effect Size

The RoBERTa-based predictions are moderate-sized effects according to Cohen (1992) and large effects according to Funder and Ozer (2019). To make sense of the effects, it is important to think broadly about the methodology: Participants described their lives to an interviewer, not being asked about their personality, and fine-tuned RoBERTa models were able to predict their personality trait levels in the testing data in the *r* = .40 range. This is perhaps a striking finding with implications for the future of personality assessment.

Convergent validity effect sizes between two self-report measures of the same construct in personality assessment research are often interpreted as high if they are above *r* = .50 (Clark & Watson, 2019). The current effect sizes approach those levels and are truly multi-method—language versus self-report. Multi-method correlations are almost always smaller than shared-method correlations (e.g., self-report versus self-report), yet the multi-method convergent validity effect sizes in the present study approach the range considered to be good in shared-method studies. This gives us the impression that *r* = .40 in this context is a large effect size.

Further, additional multi-method correlation tests between the language scores and informant-report measures of the FFM and BD are provided (Table 4). The convergent FFM language versus informant-report correlations were in the *r* = .30 range. Although there was not strong convergence between the language BD model and informant-report of BD, the language BD model did show some convergent validity with informant-reported neuroticism, low agreeableness, and low conscientiousness, which matches the BD FFM profile (Samuel & Widiger, 2008). Further benchmarks to consider in interpreting these effect sizes may be useful: Multi-method correlations between prior NLP techniques such as LIWC and self-report personality typically provide effect sizes in the .10 range. Further, the average effect size across a meta-analysis of effects in the personality literature is *r* = .21 (Fraley & Marks, 2007).

Dimensional versus Categorical model

FFM personality traits were more easily modeled by RoBERTa than PD pathology (Table 3; Figure 3). CLS embeddings indicated that people with similar FFM personality scores provided language that was more similar than people with similar PD scores. These findings indicate that linguistically, dimensional personality traits are more cohesive than PDs from the categorical model. Language, as a validation tool, supported the validity of the dimensional representation of personality traits over the categorical model of personality disorders. This finding provides reassurance that the current paradigm shift to a dimensional model of personality disorder is supported by language. Further, language as an assessment tool may be a useful way to continue to evaluate personality models.

However, the BD language model did significantly positively predict BD. And it did show some validation through convergent correlations with informant-reported FFM traits (high neuroticism, low agreeableness, low conscientiousness). Further, LIWC and BERTopic results demonstrated validity for the BD score (as well as OC and SZ). But overall, these results were significantly less robust than those for the FFM domains.

Linguistic Features of Personality and PD

Although topics provided less predictive utility, they offer unique insight into personality. Substantive topics of discussion in life narratives were important for BD and neuroticism, where topics provided rich descriptions of unique foci. Topic results provide research questions about neuroticism and BD that could improve our understanding of the development, maintenance, and consequences of these constructs, along with what may be useful coping mechanisms—for example, BD being associated with greater discussion of inner peace—a significant association with discussion of inner peace and forgiveness may reflect healing or familiarity with psychological treatment in individuals with borderline pathology in older adulthood. However, the other topics reflect significantly stressful life trials and tribulations. A discussion of finding inner peace may also be uniquely relevant to identity-based conceptualizations of PD such as the DSM-5 AMPD's Criterion A, or the general criterion of the ICD-11 model of PD.

Results may provide linguistic clues into how personality constructs may be differentiated. SZ aligns with low extraversion in research based on self-report assessment methods (Samuel & Widiger, 2008). The present study indicates that, linguistically, low extraversion is related to tentative language, differentiation, and cognitive processes. However, SZ is negatively associated with affiliation language and use of the first-person plural ("we"), whereas affiliation language is more strongly positively related to agreeableness than extraversion. These discrepancies indicate areas for future exploration, perhaps to clarify the full constellation of SZ in the FFM. Conscientiousness was negatively related to conflict language, whereas OC was positively related to conflict language. It is interesting that conscientiousness and OC are positively related to one another, but this linguistic feature is oppositely related to the two constructs. A finding like this may support that OC, as a maladaptive variant of conscientiousness, may have maladaptive behavioral manifestations at extreme levels, despite higher conscientiousness generally being a positive attribute (Carter et al., 2014, 2016; Widiger & Crego, 2019).

Implications for Research and Practice

These findings support language as a promising future direction for the assessment of personality. To our knowledge, these are the strongest language predictor models of personality from spoken language in interviews with participants to-date. It is particularly impressive that the life narrative interview does *not* ask about personality, yet the models can predict it. We conceptualize the life narrative as a proxy for an initial clinical interview (although this will have to be directly tested in the future). The initial clinical interview is a setting that will be imperative to develop language models because it is a time when a person seeking treatment is in most need of help, when a clinician is in most need of assessment help, and it would ideally not require any additional assessment time. This could be the most important advantage of language-based AI assessment in the future, as “clinicians cannot devote hours to assessment of personality disorders” (Widiger et al., 2024, p. 191), and most clinicians report they do not do formal assessment because it takes too much time (Hatfield & Ogles, 2004).

In addition, the findings have research implications. Language-based AI assessment models may provide researchers with a quick and easy multi-method assessment tool that can be implemented into any study that collects language data. Language as a multi-method assessment tool may help realize psychologists’ multi-method assessment goal (APA Task Force on Psychological Assessment and Evaluation Guidelines, 2020). Language assessment also provides unique insight into personality. The addition of novel language-based assessment tools could facilitate research progress.

Bias and Replicability

The development of AI already has a harmful history of bias against marginalized populations including Black Americans (O'Neil, 2016). The inclusion of historically marginalized groups in the development of new AI technology including language models for psychological assessment is essential. RoBERTa has shown relatively less bias compared to other LLMs regarding gender, sexuality, profession, race, and religion (Nadeem et al., 2020). A strength of the present study is the inclusion and representation of both Black and White American older adults in the dataset. However, differences across groups and explanations of those differences will be a focus of future studies. Future studies should examine model racial bias and ensure models perform equally across groups. Further, our model will perhaps only replicate with other samples of Black and White American older adults. Future research should develop models with other minority groups and continue to test models across groups.

Limitations of the Present Study

Fine-tuning of the RoBERTa language model was the most powerful prediction tool for personality in the present study, but also the most difficult to interpret. Future studies should use ways to look inside the "black box" to identify what linguistic indicators language models are using to identify personality. One way of doing this is through techniques such as visualizing attention weights, tokens, and text portions to identify which language features are most important. The addition of these techniques will be important for the further use of fine-tuning LLMs for psychological assessment because they provide more transparency and more substantive information with which to better understand personality.

Another limitation in our fine-tuning of RoBERTa is the potential of context dilution. RoBERTa has a 512-token limit. For this reason, we had to divide our transcripts into multiple “chunks.” This may cause context dilution in that each chunk may not be equally indicative of the level of the personality construct used as the label for model training (i.e., the NEO-PI-R trait domain score for an individual). We attempted to circumvent this problem by using feed forward neural networks to process the chunks. However, it is unclear how effective this was. Transformer models with longer context windows are only recently becoming more available. State-space models have also shown promise over and above transformer models and do not have the same problems with quadratic complexity with increased text length (Gu & Dao, 2023).

Finally, it will be important to train models on behavioral indicators of personality and PD in the future. Our models trained on self-report questionnaires and self-report interviews. Training models to recognize behavioral indicators such as informant-reports, behavioral tasks, real world outcomes, ambulatory assessments, and passive sensing data may provide more ecologically valid language models of personality.

Future Directions in Natural Language Processing

Language modeling is advancing at an extremely rapid pace. New LLMs show exciting possibilities for improvements in psychological assessment. For example, models with larger context windows than RoBERTa may help address the problem of context dilution. It is likely that newer and larger LLMs will increase the effect sizes of results in future studies. Compared to other advances in research methods in psychology in the recent past, the pace and possibility of advances in this area are unprecedented.

The use of language as a clinical assessment tool requires significantly more validation support. Essential areas include studies across settings and language samples, populations, assessment methods (i.e., training on behavioral data). When promising models show reliable and valid results across settings and populations, research on clinical utility will need to identify the most facilitative, appropriate, and helpful ways to implement language modeling into the clinic. In the clinic, language models can be useful tools to supplement traditional self-report assessments. Language models trained on multiple assessments may ultimately provide clinicians with scores on a variety of personality related constructs, potential outcomes, and treatment recommendations. Language models may also be trained for implementation outside the clinic through other active or passive data collection methods. This may allow better treatment progress tracking which may also facilitate treatment.

The present study focused on a community sample to identify personality constructs through language. Future studies should also examine clinical samples that have higher levels of PD pathology and examine the validity of modeling the general factor of PD from language. Additionally, we used the life narrative interview as a proxy for a traditional initial clinical interview. Although it has a similar use and structure to a clinical interview compared to other forms of language that have been used in this area previously (e.g., social media status updates), future studies should examine actual clinical histories/interviews to ensure the transferability of models developed in life narrative interviews to clinical histories.

Language is a robust predictor of psychological constructs. However, other features such as speech acoustics and facial features may add predictive utility to models of PD as well. Although it is likely that language is the strongest predictor of the three, it is true that certain speech acoustic or facial/movement features may be extremely important for predicting nuanced features of PD. For example, speech rate is an especially important predictor of depression (Cummins et al., 2015).

Finally, the present study focused on the FFM traits and three DSM PDs. Although the current dimensional personality trait representations are based on the FFM, findings here reinforce the idea that pathological traits are different than normal range traits and it will be important to train models on modern assessments of dimensional maladaptive personality traits in the DSM-5 and ICD-11. Additionally, it will be imperative to train models on general severity of personality pathology such as those included in the DSM-5 AMPD and ICD-11 general severity criterion. Not until we have language modeling in each area (i.e., dimensional traits, general personality functioning, and traditional PD categories) will we be able identify strengths and limitations across models and use language as a PD assessment tool to advance research, knowledge, and ultimately clinical practice.

Conclusions

Language-based AI assessment of personality has exciting potential to advance PD research and practice. The results of the present study provide clear evidence of those possibilities. Despite limitations, the implications of these findings are somewhat remarkable: From a broad life narrative interview, with no instructions to discuss personality, language models can reproduce self-reported personality scores that demonstrate multi-method construct validity support. In sum, we believe further research developing AI for personality assessment could contribute significantly to the advancement of substantive research knowledge and the clinical treatment of people with PD.

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

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck depression inventory-II*. The Psychological Corporation.

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., M. Gupta, & Oltmanns, J. R. (2025). Large Language Models for Psychological Assessment: A Comprehensive Overview. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/qm9ae_v1

Calabrese, W. R., Emery, L. T., Evans, C. M., & Simms, L. J. (2024). Diagnostic and Statistical Manual of Mental Disorders, fifth edition, personality disorders and the alternative model: Prediction of naturalistically observed behavior, interpersonal functioning, and psychiatric symptoms, 1 year later. *Personality Disorders: Theory, Research, and Treatment*, *15*(5), 361–370. https://doi.org/10.1037/per0000677

Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M.-C., & Delgado, K. M. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes an empirical difference. *Journal of Applied Psychology*, *99*(4), 564–586. https://doi.org/10.1037/a0034688

Carter, N. T., Guan, L., Maples, J. L., Williamson, R. L., & Miller, J. D. (2016). The Downsides of Extreme Conscientiousness for Psychological Well-being: The Role of Obsessive Compulsive Tendencies. *Journal of Personality*, *84*(4), 510–522. https://doi.org/10.1111/jopy.12177

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

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

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

Cruitt, P. J., & Oltmanns, T. F. (2018). Incremental Validity of Self- and Informant Report of Personality Disorders in Later Life. *Assessment*, *25*(3), 324–335. https://doi.org/10.1177/1073191117706020

Cruitt, P. J., & Oltmanns, T. F. (2019). Unemployment and the relationship between borderline personality pathology and health. *Journal of Research in Personality*, *82*, 103863. https://doi.org/10.1016/j.jrp.2019.103863

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49. https://doi.org/10.1016/j.specom.2015.03.004

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

Digman, J. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417–440.

Entwistle, C., Horn, A. B., Meier, T., Hoemann, K., Miano, A., & Boyd, R. L. (2023). Natural emotion vocabularies and borderline personality disorder. *Journal of Affective Disorders Reports*, *14*, 100647. https://doi.org/10.1016/j.jadr.2023.100647

Fraley, R. C., & Marks, M. J. (2007). The Null Hypothesis Significance-Testing Debate and Its Implications for Personality Research. In *Handbook of research methods in personality psychology* (pp. 149–169). Guilford Press.

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

Gleason, M. E. J., Powers, A. D., & Oltmanns, T. F. (2012). The enduring impact of borderline personality pathology: Risk for threatening life events in later middle-age. *Journal of Abnormal Psychology*, *121*(2), 447–457. https://doi.org/10.1037/a0025564

Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. U. California 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

Gu, A., & Dao, T. (2023). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. arXiv preprint arXiv:2312.00752.

Hatfield, D. R., & Ogles, B. M. (2004). The Use of Outcome Measures by Psychologists in Clinical Practice. *Professional Psychology: Research and Practice*, *35*(5), 485–491. https://doi.org/10.1037/0735-7028.35.5.485

Hays, R. D., Prince-Emburg, S., & Chen, H. (1998). *Rand-36 Health Status Inventory*. The Psychological Corporation.

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C. P., Deters, F. G., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., & Mehl, M. R. (2019). Linguistic Markers of Grandiose Narcissism: A LIWC Analysis of 15 Samples. *Journal of Language and Social Psychology*, *38*(5–6), 773–786. https://doi.org/10.1177/0261927X19871084

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). *From Text to Thought: How Analyzing Language Can Advance Psychological Science*.

Jain, D., Arora, S., Jha, C. K., & Malik, G. (2024). Text classification models for personality disorders identification. *Social Network Analysis and Mining*, *14*(1), 64. https://doi.org/10.1007/s13278-024-01219-8

Jain, D., Kumar, A., & Beniwal, R. (2022). Personality BERT: A Transformer-Based Model for Personality Detection from Textual Data. In A. K. Bashir, G. Fortino, A. Khanna, & D. Gupta (Eds.), *Proceedings of International Conference on Computing and Communication Networks* (Vol. 394, pp. 515–522). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-0604-6_48

Kish, L. (1949). A Procedure for Objective Respondent Selection within the Household. *Journal of the American Statistical Association*, *44*(247), 380–387.

Lawson, M. A., & Matz, S. C. (2022). Saying more than we know: How language provides a window into the human psyche. In S. C. Matz (Ed.), *The psychology of technology: Social science research in the age of Big Data.* (pp. 45–85). American Psychological Association. https://doi.org/10.1037/0000290-003

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

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*(5), 862–877. https://doi.org/10.1037/0022-3514.90.5.862

Mehta, Y., Fatehi, S., Kazameini, A., Stachl, C., Cambria, E., & Eetemadi, S. (2020). Bottom-Up and Top-Down: Predicting Personality with Psycholinguistic and Language Model Features. *2020 IEEE International Conference on Data Mining (ICDM)*, 1184–1189. https://doi.org/10.1109/ICDM50108.2020.00146

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (No. arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Nadeem, M., Bethke, A., & Reddy, S. (2020). *StereoSet: Measuring stereotypical bias in pretrained language models* (No. arXiv:2004.09456). arXiv. http://arxiv.org/abs/2004.09456

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., Schwartz, H. A., Ruggero, C., Son, Y., Miao, J., Waszczuk, M., Clouston, S. A. P., Bromet, E. J., Luft, B. J., & Kotov, R. (2021). Artificial intelligence language predictors of two-year trauma-related outcomes. *Journal of Psychiatric Research*, *143*, 239–245. https://doi.org/10.1016/j.jpsychires.2021.09.015

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

Pennebaker, J. W., & King, L. A. (1999). Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations* (No. arXiv:1802.05365). arXiv. http://arxiv.org/abs/1802.05365

Pfohl, B., Blum, N., & Zimmerman, M. (1997). *Structured interview for DSM-IV personality: SIDP-IV*. American Psychiatric Association Publishing.

Powers, A. D., Gleason, M. E. J., & Oltmanns, T. F. (2013). Symptoms of borderline personality disorder predict interpersonal (but not independent) stressful life events in a community sample of older adults. *Journal of Abnormal Psychology*, *122*(2), 469–474. https://doi.org/10.1037/a0032363

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

Samuel, D., & Widiger, T. (2008). A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: A facet level analysis☆. *Clinical Psychology Review*, *28*(8), 1326–1342. https://doi.org/10.1016/j.cpr.2008.07.002

Sanford, F. H. (1942). Speech and personality. *Psychological Bulletin*, *39*(10), 811–845.

Simchon, A., Sutton, A., Edwards, M., & Lewandowsky, S. (2023). Online reading habits can reveal personality traits: Towards detecting psychological microtargeting. *PNAS Nexus*, *2*(6), pgad191. https://doi.org/10.1093/pnasnexus/pgad191

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

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvy, D. M. (1966). *The General Inquirer: A computer approach to content analysis*. MIT Press.

Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., Edwards, T. S., Pennebaker, J. W., & Mehl, M. R. (2019). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of Personality and Social Psychology*, *116*(5), 817–834. https://doi.org/10.1037/pspp0000187

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *31st Conference on Neural Information Processing Systems*, 11.

Weissman, M. M. (1999). *SAS-SR Question Booklet*. Multi Health Systems.

Weizenbaum, J. (1961). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Widiger, T. A., & Crego, C. (2019). The bipolarity of normal and abnormal personality structure: Implications for assessment. *Psychological Assessment*, *31*(4), 420–431. https://doi.org/10.1037/pas0000546

Widiger, T. A., Hines, A., & Crego, C. (2024). Evidence-Based Assessment of Personality Disorder. *Assessment*, *31*(1), 191–198. https://doi.org/10.1177/10731911231176461

World Health Organization. (2025). *International statistical classification of diseases (11th ed.)*. https://icd.who.int/

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

Xu, B., Wang, N., Chen, T., & Li, M. (2015). *Empirical Evaluation of Rectified Activations in Convolutional Network* (No. arXiv:1505.00853). arXiv. http://arxiv.org/abs/1505.00853

Table 1. *Participant Demographics*

Variable	*N*	Percent	Mean	*SD*
Age	1409		59.5	3
Gender				
Female	771	45.3%		
Male	638	54.7%		
Race				
White	916	65.0%		
Black/African American	460	32.6%		
Non-Black Latino	11	0.8%		
Biracial/Multiracial	7	0.5%		
Middle Eastern	4	0.4%		
Other	10	0.8%		
Marital Status				
Married	679	48.2%		
Divorced	406	28.8%		
Never Married	200	14.2%		
Widowed	98	7.0%		
Separated	26	1.8%		
Education				
H.S. Diploma	381	27.0%		
Bachelor Degree	359	25.5%		
Master Degree	265	18.8%		
Associate Degree	132	9.4%		
Doctorate	106	7.5%		
Vocational Tech Degree	75	5.3%		
Elementary or Junior High	34	2.4%		
GED	31	2.2%		
R. N. Diploma	23	1.6%		
Don't Know	3	0.2%		
Annual Household Income				
Under $20,000	168	11.9%		
$20,000-$39,999	247	17.5%		
$40,000-$59,999	292	20.7%		
$60,000-$79,999	177	12.6%		
$80,000-$99,999	133	9.4%		
$100,000-$119,999	99	7.0%		
$120,000-$139,999	63	4.5%		
$140,000 or more	165	11.7%		
Missing	65	4.6%		

Table 2

Descriptive Statistics for the NEO-PI-R and SIDP-IV Personality Scores

Personality Score	Mean	*SD*	Skewness	Minimum	Maximum	Median
Neuroticism	72.83	21.46	0.52	13.00	162.00	71.00
Extraversion	109.70	19.34	0.07	47.00	216.00	110.00
Openness	113.76	19.29	0.46	51.00	211.50	113.00
Agreeableness	131.25	16.61	0.32	66.00	220.50	131.00
Conscientiousness	125.05	19.14	0.53	39.00	241.50	125.00
SZ	0.28	0.68	3.46	0.00	6.00	0.00
BP	0.21	0.64	4.58	0.00	7.00	0.00
OC	0.71	1.06	1.85	0.00	7.00	0.00
	Symptom Frequencies					
	1	2	3	4	5	6+
SZ	195	50	16	6	2	2
BD	145	35	13	4	2	3
OC	331	166	53	26	9	4

Note. OC = obsessive-compulsive.

Table 3

RoBERTa, BERTopic, LIWC, and Combined Language Modeling of Personality

	Neuroticism		Extraversion		Openness		Agreeableness		Conscientious		Borderline		Schizoid		Obsessive-Compulsive	
	MSE	*r*	*MSE*	*r*	*MSE*	*r*	*MSE*	*r*	*MSE*	*r*	*MSE*	*r*	*MSE*	*r*	*MSE*	*r*
LIWC	.94	.24	.96	.20	.92	.26	.90	.26	-		.96	.17	.96	.14	.97	.14
BERTopic	.96	.20	.97	.17	.97	.17	.97	.17	.94	.22	.27	.14	.35	.14	.38	.10
RoBERTa	.81	**.42**	.84	**.39**	.79	**.45**	.89	**.32**	.89	**.32**	.22	.26	-		-	
Combined	.82	**.41**	.93	**.41**	.80	**.45**	.89	**.32**	.92	.26	.89	**.35**	1.05	.14	1.05	.14

Note. Moderate *r* effect sizes in bold. Combined model N, E, O, A, C, & BP *n*'s = 1,013. SZ and OC *n*'s = 1,054.

Table 4

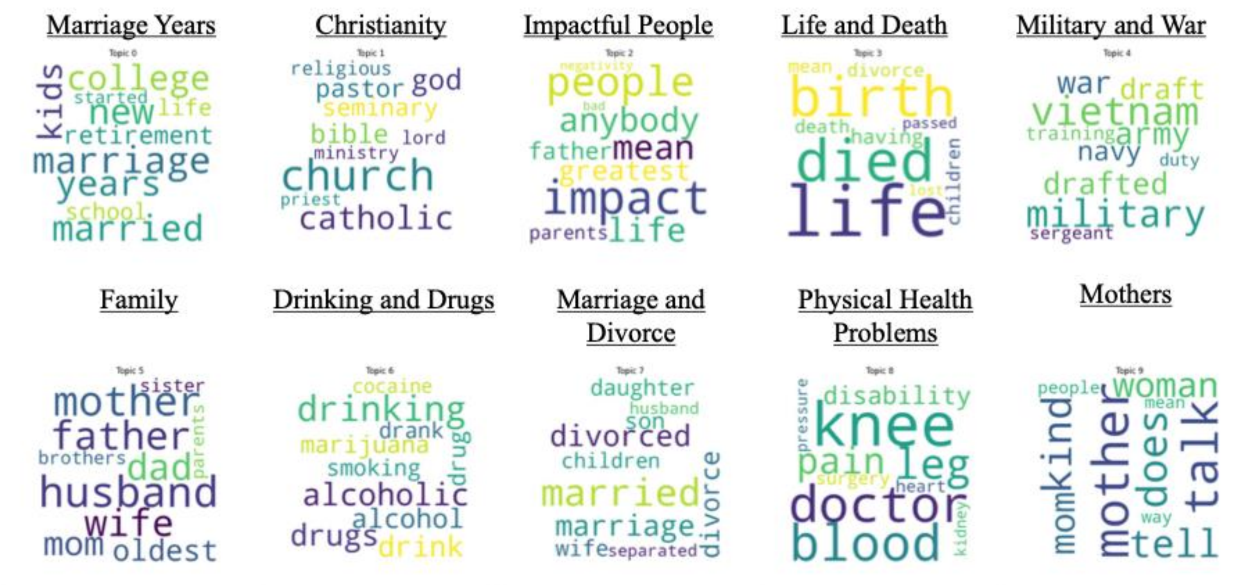
Correlations Between Fine-Tuned RoBERTa Language Models' Test Set Predictions and Multimethod Criteria

						Inf-	Self-	Self-	Inf-	Inf-Soc
Fine-Tuned RoBERTa	Inf-Neur	Inf-Ext	Inf-Ope	Inf-Agre	Inf-Con	Bor	PF	Dep	Health	Adj
Neuroticism	**.36** ***	-.15 *	-.07	-.09	-.31 ***	.14 *	-.21 **	.30 ***	.26 ***	.28 ***
Extraversion	-.17 *	**.35** ***	.17 *	.03	.11	-.06	.18 **	-.16 *	-.17 *	-.19 **
Openness	.08	.05	**.31** ***	-.04	.05	.04	.16 **	-.02	-.04	.04
Agreeableness	.09	.14	.07	**.26** ***	.09	.03	.01	-.07	-.03	-.13
Conscientiousness	-.25 ***	.19 **	.10	.02	**.28** ***	-.03	.14 *	-.19 **	-.15 *	-.26 ***
Borderline	.22 ***	-.07	-.04	-.24 ***	-.31 ***	**.11**	-.11	.15 *	.24 ***	.22 **

Note: * = *p* < .05; *** = *p* < .01, *** = *p* < .001. Convergent multi-method correlations in bold. Inf = informant-report, Neur = neuroticism, Ext= extraversion, Ope = openness, Agr = agreeableness, Con = conscientiousness, Bor = borderline, Self = self-report, PF = physical functioning, Dep = depressive symptoms, Health = physical and emotional health status, Soc Adj = social adjustment. *n's* range from 206 (informant openness) to 266 (physical functioning).

Figure 1

Top 10 most Prevalent Topics in the Life Narrative Interviews.



Marriage Years

Christianity

Impactful People

Life and Death

Military and War

Family

Drinking and Drugs

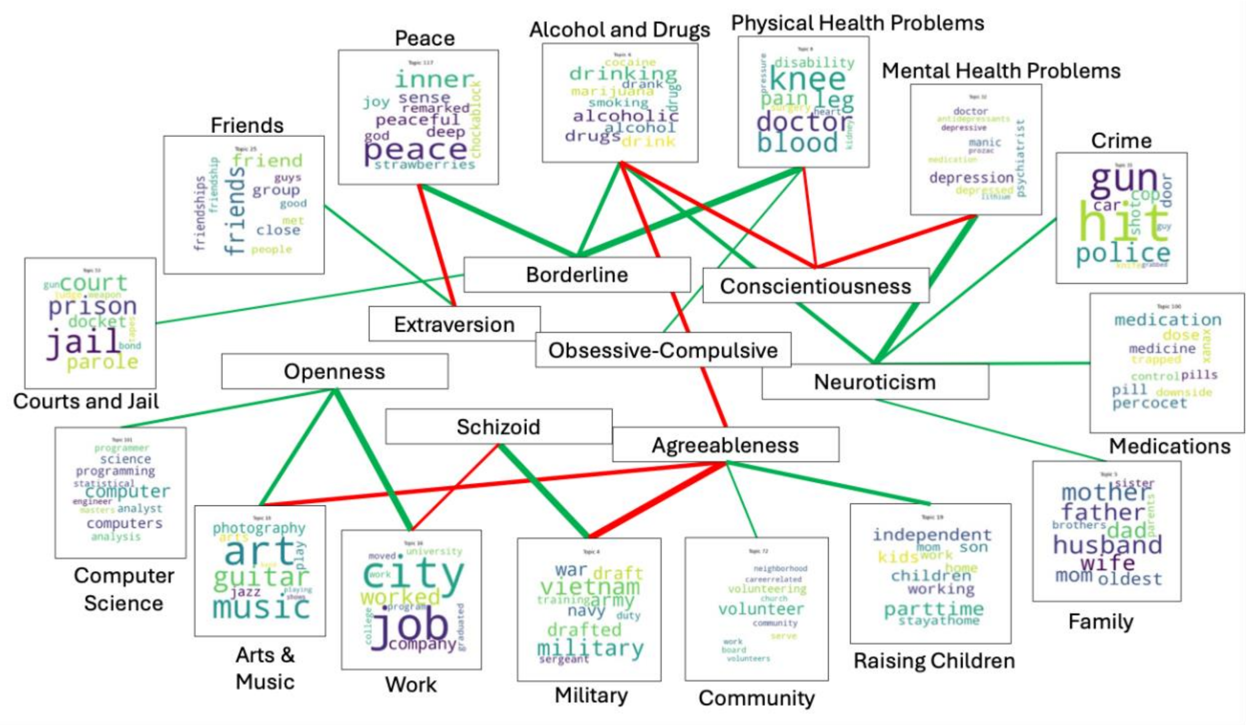
Marriage and Divorce

Physical Health Problems

Mothers

Figure 2

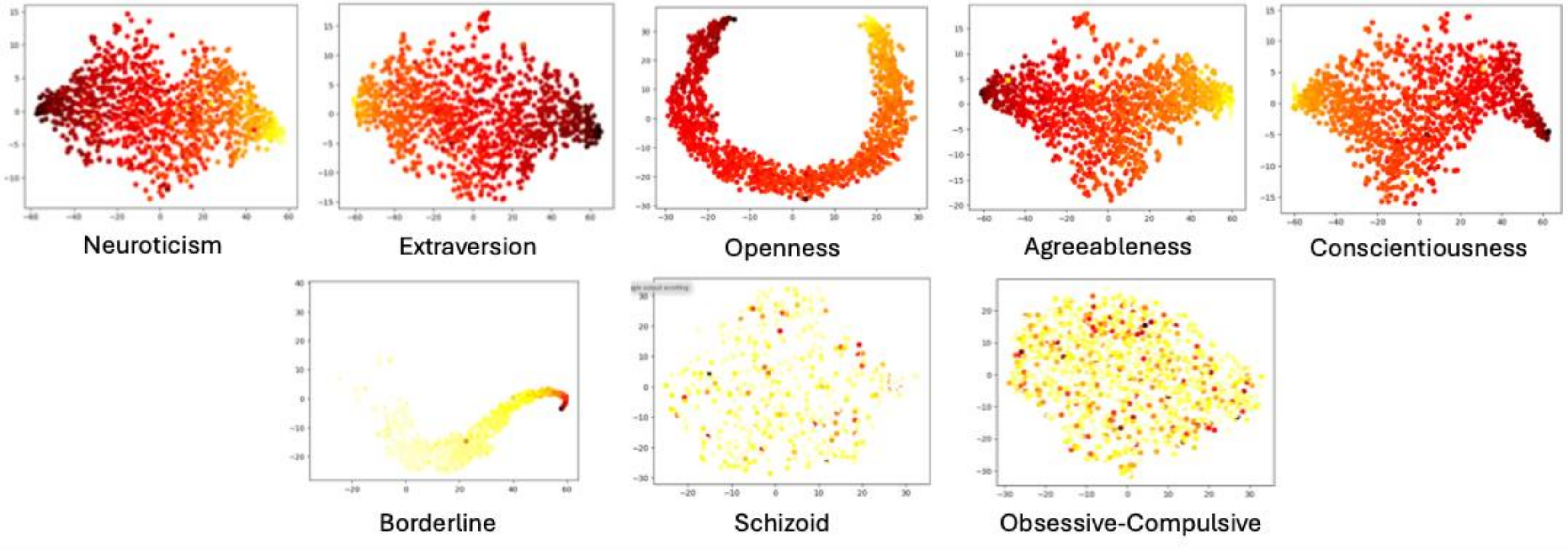
Topics Correlated with Personality



Note. Significant correlations are indicated by connection lines and range from *r* = .05 to *r* = .14. Relatively stronger correlations are indicated by thicker connection lines.

Figure 3

2D Representation of the CLS Embeddings for the Personality Scores from the Life Narrative Texts



Neuroticism

Extraversion

Openness

Agreeableness

Conscientiousness

Borderline

Schizoid

Obsessive-Compulsive