

Original Paper

Personality Inference From Semistructured Interviews Using Large Language Models: Comparative Evaluation Study

Jianfeng Zhu¹, PhD; Julina Maharjan¹, PhD; Xinyu Li¹; Karin G Coifman², Prof Dr; Ruoming Jin¹, Prof Dr

¹Department of Computer Science, Kent State University, Kent, OH, United States

²Department of Psychological Sciences, Kent State University, Kent, OH, United States

Corresponding Author:

Jianfeng Zhu, PhD
Department of Computer Science
Kent State University
241 Mathematical Sciences Building
Kent, OH 44242
United States
Phone: 1 3306729107
Email: jzhu10@kent.edu

Abstract

Background: Personality traits are central to understanding human behavior and life outcomes, yet their assessment remains challenging due to biases and limited expressiveness in traditional self-report inventories. The emergence of large language models (LLMs) creates new opportunities to infer personality from naturalistic text, but it remains unclear whether these systems can reliably predict validated personality traits or whether their predictions primarily reflect superficial linguistic patterns.

Objective: This study evaluates the ability of LLM-based methods to predict continuous Big Five personality scores (1-5 Likert scale) from semistructured interview transcripts and compares the performance of different modeling paradigms.

Methods: We analyzed semistructured interview transcripts from 518 participants, each paired with validated self-reported Big Five scores measured using the 10-item Big Five Inventory (BFI-10). We compared four paradigms: (1) zero-shot and chain-of-thought prompting with GPT-5 mini; (2) parameter-efficient fine-tuning (low-rank adaptation) applied to open-source RoBERTa-base and LLaMA-2-7B models; (3) supervised fine-tuning of GPT-4.1 mini across two independent training runs; and (4) embedding-based regression using 4 pretrained encoders (all-mpnet-base-v2, E5-large-v2, text-embedding-3-small, and text-embedding-3-large) with a multioutput ridge regression.

Results: Supervised fine-tuning yielded the strongest overall performance among the evaluated approaches. In the best-performing run, correlations reached $r=0.718$ for conscientiousness and $r=0.631$ for agreeableness. Neuroticism showed moderate and consistent performance across the 2 fine-tuning runs, whereas extraversion was less stable across runs. Embedding-based regression, particularly with text-embedding-3-large, achieved competitive results for conscientiousness and agreeableness, although it remained below supervised fine-tuning. Prompt-based methods showed weak and inconsistent performance across traits. Openness was generally difficult to predict, although RoBERTa-base with embedding-based low-rank adaptation fine-tuning showed modest but statistically significant performance, indicating trait- and method-dependent variability.

Conclusions: LLM-based approaches exhibit trait-dependent performance in predicting personality from naturalistic dialogue, with fine-tuning substantially improving correspondence with validated psychometric scores. These findings suggest that personality inference from naturalistic dialogue is feasible but depends strongly on modeling strategy, trait domain, and the availability of supervised fine-tuning with validated personality measures.

J Med Internet Res 2026;28:e87129; doi: [10.2196/87129](https://doi.org/10.2196/87129)

Keywords: large language models; LLMs; personality inference; Big Five personality traits; semistructured interviews; low-rank adaptation; LoRA fine-tuning

Introduction

Background and Study Rationale

Personality traits play a fundamental role in understanding individual differences, social behavior, and life outcomes [1-3]. However, their assessment remains challenging, as traditional self-report inventories are subject to response biases and may be limited in their ability to capture the richness of individual expression [4,5]. Prior research suggests that personality-relevant information is reflected in everyday language use, as individuals express stable patterns of thought and behavior in how they speak and write [6-8]. This perspective is grounded in the lexical hypothesis, which posits that meaningful aspects of personality are encoded in natural language and expressed through everyday communication [9-11].

Building on this foundation, recent advances in large language models (LLMs) have enabled personality prediction directly from text, offering scalable and context-sensitive alternatives to traditional assessment methods [12-18]. Existing studies suggest that LLMs can infer Big Five traits from language with modest accuracy, primarily using social media data or written corpora [19-21]. Approaches leveraging extended conversational data, such as interview transcripts, have shown improved performance, suggesting that richer, context-aware language better captures personality-relevant signals [16,22-24].

However, existing work remains limited in 3 key respects. First, many studies frame personality as categorical labels rather than continuous traits, limiting psychometric fidelity [14,25]. Second, most datasets rely on social media text [26-29], which may not reflect structured or task-oriented conversational contexts. Third, few studies jointly integrate ecologically grounded interview data [30], validated individual-level trait measures, and systematic comparisons of multiple LLM modeling paradigms under a unified evaluation framework. As a result, it remains unclear how well LLMs can predict personality from interview data in ways that align with established psychometric standards. Our prior research began to explore this challenge using prompt-based inference on the same interview-based benchmark [24], and the present study extends that work by incorporating supervised fine-tuning and representation-based regression approaches.

The Big Five, or Five-Factor Model, is the most widely accepted framework in personality psychology [31], comprising 5 broad trait dimensions: openness, conscientiousness, agreeableness, extraversion, and neuroticism. To address the gaps identified above, this study evaluates whether contemporary LLMs can predict continuous Big Five personality scores from semistructured interview transcripts paired with validated 10-item Big Five Inventory (BFI-10) self-report scores. This design better approximates naturalistic human communication and enables a more realistic assessment of model performance against psychometrically grounded benchmarks.

We compare 4 complementary modeling paradigms reflecting distinct assumptions about how personality signals are encoded and extracted from text. First, we examine prompt-based inference using GPT-5 mini (OpenAI) under both zero-shot and chain-of-thought (CoT) settings, representing inference without task-specific adaptation. Second, we evaluate parameter-efficient fine-tuning via low-rank adaptation (LoRA) applied to RoBERTa-base (Robustly Optimized BERT Approach; Meta AI) and LLaMA-2-7B (Large Language Model Meta AI; Meta Platforms, Inc.). Third, we investigate supervised model adaptation through supervised fine-tuning of GPT-4.1 mini (OpenAI), which we evaluate across 2 independent training runs. Fourth, we assess embedding-based regression, in which fixed representations from 4 pretrained encoders (all-mpnet-base-v2 [Microsoft], E5-large-v2 [Microsoft], text-embedding-3-small [OpenAI], and text-embedding-3-large [OpenAI]) are used to predict personality traits via a multioutput ridge regression model. By contrasting these paradigms, we aim to disentangle the contributions of prompting strategy, representation learning, and different levels of model adaptation in personality inference and to identify which traits are most reliably captured from naturalistic conversational data.

Related Work

Traditional Text-Based Personality Recognition

Automatic personality recognition has long been studied in natural language processing. Early approaches relied on classical machine learning techniques (eg, support vector machines and regression) [32], combined with handcrafted linguistic features derived from resources such as linguistic inquiry and word count categories [33], emotion lexicons [34], and term frequency-inverse gravity moment [35]. Applied to essays, blogs, and social media text, these methods demonstrated that stable personality signals are present in written language [36]. Meta-analytic evidence supports this pattern, although effect sizes typically remain modest and vary considerably across traits, platforms, and modeling choices [37,38].

Subsequent advances introduced transformer-based models such as BERT (bidirectional encoder representations from transformers) [39,40], which leverage large-scale pretraining and fine-tuning on datasets such as myPersonality [41] and PANDORA [42]. These models improved predictive accuracy but continued to rely on binary or categorical labels and nonclinical data, limiting their psychometric fidelity and generalizability. These limitations have motivated the exploration of more flexible, language-driven approaches based on LLMs.

Data Paradigms: Social Media, Author Profiling, and Conversational Data

Much of the existing work in personality recognition is grounded in social media and author profiling datasets. Shared tasks such as the PAN 2015 Author Profiling task and datasets such as PANDORA have provided valuable scale for personality computing, but they are largely grounded in tweets or Reddit posts rather than elicited, topic-diverse

narratives [25,28,29,41,43,44]. However, these datasets are primarily based on short-form text (eg, tweets or Reddit posts), which may not reflect how individuals express themselves in structured or context-rich settings.

In parallel, persona-conditioned dialogue resources such as Persona-Chat [45] have been developed to support controllable text generation conditioned on predefined personality profiles. While useful for generation tasks, such datasets do not capture validated trait scores from real individuals and, therefore, are not directly suitable for psychometric evaluation.

LLM-Based Personality Inference and Modeling Paradigms

Recent studies have examined LLMs for personality-related language tasks, ranging from personality-conditioned dialogue generation to personality trait inference. Prompt-based methods, including zero-shot, few-shot prompting, and CoT prompting, provide lightweight strategies for eliciting personality-related judgments without task-specific fine-tuning [24,46-48]. Peters and Matz [29] reported that GPT-3.5 and GPT-4 achieved moderate correlations with self-reported traits using zero-shot prompting on Facebook (Meta Platforms, Inc) status updates. Similarly, Piastra and Catellani [19] found small-to-moderate predictive performance when applying ChatGPT (OpenAI) to written text. More structured approaches, such as PsyCoT, use psychological questionnaire items as explicit reasoning steps and have been shown

to improve GPT-3.5 (OpenAI) performance on personality detection benchmarks, although these gains have primarily been demonstrated in classification settings rather than in continuous trait prediction [49].

More substantial progress has been achieved through approaches that incorporate supervised learning and parameter adaptation. Embedding-based methods, which use LLM-derived representations as inputs to downstream models, have been shown to outperform prompt-based inference in several settings [25]. Parameter-efficient fine-tuning techniques, such as LoRA [50], further improve adaptation while reducing computational cost, enabling scalable personality prediction across domains [51]. Recent frameworks such as PersLLM [52] extend this idea by combining LLM-based representation learning with lightweight output networks and memory layers, reducing redundant computation while maintaining competitive performance. In parallel, retrieval-augmented methods, including retrieval-augmented generation (RAG) [53] and PersonaRAG [54], introduce mechanisms for grounding predictions in external or user-specific data, highlighting the potential of personalized and context-aware inference.

Table 1 summarizes major LLM-based personality recognition studies published between 2021 and 2025, highlighting rapid methodological innovation alongside persistent challenges in data quality, labeling schemes, and interpretability.

Table 1. Comparative overview of recent large language model (LLM)-based personality recognition studies published between 2021 and 2025.

Study	LLM(s) used	Methodology	Dataset	Label type	Key finding/limitation
Christian et al [55], 2021	BERT, RoBERTa, and XLNet	Used a multimodel deep learning architecture using pretrained language models as feature extractors, combined with model averaging	myPersonality	Binary	Outperformed previous works by combining multiple data sources and models, achieving high classification accuracy (eg, 88.5% on Twitter [subsequently rebranded X] data)
Ji et al [56], 2023	ChatGPT	Used prompting strategies (zero-shot, one-shot, and CoT ^a); zero-shot CoT most effective for explanation and prediction	Essays and PANDORA	Binary	CoT improves over basic zero-shot but still trails fine-tuned state-of-the-art models
Sirasapalli and Malla [57], 2023	BERT	Proposed data fusion, mapping MBTI ^b dataset to Big Five traits, and combining with other corpora	Essays, myPersonality, and MBTI	Binary	Data fusion increased labeled data and model performance, reaching 87.9% accuracy
Serapio-García et al [15], 2025	PaLM family (18 LLMs)	Assessed LLM personality using validated psychometric tests (IPIP-NEO ^c and BFI ^d) under various prompting	Psychometric Tests	N/A ^e (assessment)	Simulated LLM personality was reliable and valid; shaped via prompting
Hu et al [58], 2024	LLM (unspecified)	Used LLMs to generate augmentations for contrastive distillation	Kaggle personalitycafe and PANDORA	Binary	LLM-augmented data improves task-specific models
Peters and Matz [29], 2024	GPT-4, LLaMA 2, and Mistral	Used zero-shot and CoT for continuous prediction; assessed with correlation and error	Essays	Continuous	Weak correlation ($r<0.30$); minimal gain from CoT

Study	LLM(s) used	Methodology	Dataset	Label type	Key finding/limitation
Wang et al [59], 2025	GPT-4	Emulated individuals using personality scores; completed questionnaires as that in-person	Human personality data	Continuous	Emulated responses had high convergent validity ($r>0.90$) but superior (less realistic) consistency
Maharjan et al [25], 2025	RoBERTa, BERT, and OpenAI	Generated LLM embeddings and trained bidirectional long short-term memory model for continuous scores	PANDORA	Continuous	Embedding models outperform zero-shot
This work	GPT-4.1 mini, GPT-5 mini, LLaMA-2-7B, and RoBERTa-base	Compares zero-shot and CoT prompting, LoRA ^f fine-tuning, supervised fine-tuning, and embedding-based regression on interview transcripts, with validated Big Five scores	Real-world interviews	Continuous	Provide a systematic comparison of 4 LLM paradigms on ecologically grounded interview data with validated psychometric labels; fine-tuning substantially outperforms prompting

^aCoT: chain-of-thought.

^bMBTI: Myers-Briggs Type Indicator.

^cIPIP-NEO: International Personality Item Pool–neuroticism, extraversion, and openness.

^dBFI: Big Five Inventory.

^eN/A: not applicable.

^fLoRA: low-rank adaptation.

Collectively, prior work highlights both the promise and the fragmentation of LLM-based personality inference, with substantial variation in data sources, labeling schemes, and modeling paradigms. Our prior work evaluated LLMs (eg, GPT-5 mini, GPT-4.1 mini, Meta-LLaMA [Meta Platforms, Inc], and DeepSeek) using zero-shot and CoT prompting on the same interview-based benchmark and observed consistently weak performance in personality prediction [24].

These findings suggest that limitations arise not only from model capacity but also from reliance on prompt-based inference. The present study moves beyond prompt-based inference by introducing supervised fine-tuning and representation-based approaches, including model fine-tuning and embedding-based regression, to enable explicit learning from labeled data.

Dataset and Participants

In this study, we used real-world, semistructured interview transcripts from a pooled dataset of 555 adults across 5 prior investigations of adjustment to stress. The pooled source dataset included active-duty firefighters ($n=103$), patients with multiple sclerosis ($n=61$), adults recruited following injury or traumatic events ($n=254$), first-year college students ($n=84$), and community adults ($n=53$). Of the 555 participants in the pooled source dataset, 518 had complete BFI-10 responses and were included in the present analysis. The mean scores for the Big Five traits are as follows: openness (3.35, SD 0.63), extraversion (3.03, SD 1.09), neuroticism (2.93, SD 1.23), conscientiousness (2.74, SD 1.30), and agreeableness (2.63, SD 1.13), indicating that average trait levels were near the midpoint of the scale. Full descriptive statistics are reported in Table A7 of [Multimedia Appendix 1](#).

The analytic sample was demographically diverse and balanced by sex ($n=260$ male and $n=257$ female participants), had minimal missing data, and spanned a wide age range (mean 39.32, SD 16.45). Participants represented various racial backgrounds ($n=402$ White and $n=116$ non-White/Other participants) and educational levels ranging from high school to college and above. Additional details regarding cohort characteristics and interview prompts are available via the Open Science Framework (OSF; Center for Open Science) repository [60]. The complete semistructured interview prompts are provided in Table A1 of [Multimedia Appendix 1](#). All code used in this study is publicly available at the project repository [61].

The semistructured interview protocol varied across the 5 parent studies to reflect the specific context and circumstances of each sample. Depending on the parent study, participants completed 3 to 5 prompts, beginning with a neutral description of previous-day activities and followed by study-specific prompts concerning relevant experiences, coping, and/or positive and negative events. Participants were encouraged to speak spontaneously for up to 3 minutes in response to each prompt [62–64]. Interviews were audio-recorded and transcribed using established conventions. The resulting transcripts showed substantial variability in length. On average, participants produced 2955 (SD 1855) words per interview, with a range from 387 to 7143 words (median 2292, IQR 1569–4388). Representative excerpts from 1 participant in the active-duty firefighter parent study are presented in [Table 2](#).

Table 2. Representative responses from a semistructured experiential interview with 1 participant from the active-duty firefighter parent study, selected from the pooled analytic sample (N=518).

Administered prompt	Response excerpt
Baseline: previous-day activities	“Beginning of the day... I have two sons... played outside... gym...”
Q1: experience of being a firefighter	“Being a firefighter... challenging and amazing experience... bad experiences too...”
Q2: coping with firefighter work	“You talk to people you trust at work... my wife and I have been married...”
Q3: scary or frightening event	“First big incident... stressful leading into it...”
Q4: positive firefighter event	“First baby we delivered—early morning call... heroin case...”

Methods

Prompt-Based Inference

To evaluate the capacity of LLMs for personality inference, we used the GPT-5 mini via the OpenAI API (model ID: *gpt-5-mini-2025-08-07*; [65]) to predict Big Five trait scores from naturalistic interview transcripts. The model was instructed to act as a psychologically insightful agent and to generate continuous scores between 1.0 and 5.0 for each of the five traits.

We implemented 2 prompting strategies. The first was a standard zero-shot prompt that directly instructed the model to return trait scores based on the interview content [66]. The second was a CoT prompt that encouraged intermediate reasoning about emotional tone and linguistic markers prior to scoring [67]. For prompt-based inference, transcripts were used in their conversational form, with interviewer prompts preserved to retain contextual structure. Both prompting strategies were executed using the same API configuration without specifying a temperature parameter. Model outputs followed a predefined score format, and trait scores were automatically extracted using rule-based parsing implemented with regular expressions. Compared with our prior work [24], the prompts were slightly refined with additional instructions while maintaining the same overall prompting structure. The complete zero-shot and CoT prompt templates, along with an example model output, are provided in the Textboxes 1 to 3 of [Multimedia Appendix 1](#).

LoRA-Based Adaptation

To improve prediction performance with respect to self-reported Big Five personality scores, we applied LoRA [50] for efficient fine-tuning of transformer-based language models. Here, *r* denotes the rank of the low-rank decomposition, representing the dimensionality of the trainable adaptation matrices. A larger rank increases model capacity but requires more parameters, whereas a smaller rank improves parameter efficiency. Specifically, we fine-tuned the open-source LLaMA-2-7B model [68] to jointly predict all 5 personality traits from the input text. LoRA modules are injected into the attention and feedforward layers, allowing for parameter-efficient adaptation while preserving the model’s general language capabilities.

Our implementation used LoRA fine-tuning with 4-bit quantization to reduce memory requirements. The training configuration used a learning rate of 1e-4, a per-device batch

size of 2, gradient accumulation steps of 4, weight decay of 0.01, and 3 training epochs. The LoRA configuration used a scaling factor α of 64, a dropout rate of 0.05, and no bias terms. In the rank-sensitivity experiments, adapter ranks of 8, 16, and 32 were evaluated using the same configuration unless otherwise specified.

In addition to decoder-based models, we evaluated encoder-only architectures to assess the role of representation learning in personality prediction. We used RoBERTa-base [39] as a representative encoder model. Because of its limited input length, interview transcripts were tokenized and segmented into nonoverlapping 512-token chunks, with an overlap ratio of 0.0 and a stride of 512 tokens. Each chunk was encoded independently, and chunk-level embeddings were aggregated using mean pooling to form a global transcript representation. A lightweight nonlinear regression head was then trained on these pooled embeddings to predict continuous trait scores.

Supervised Fine-Tuning of GPT-4.1 mini

To examine whether supervised model adaptation improves personality trait prediction, we conducted supervised fine-tuning using a pretrained, closed-source LLM. Specifically, we used GPT-4.1 mini (April 14, 2025; [69]) as the base model and trained it on the personality prediction dataset.

Each training instance consisted of a participant’s textual responses as the input and the corresponding continuous scores for the 5 Big Five personality traits as the target outputs. The dataset was split into an 80% fine-tuning set and a 20% test set. Fine-tuning was performed using a supervised learning objective for 3 epochs, with a batch size of 1 and a learning rate multiplier of 2. No separate validation set was used.

To assess the consistency of the fine-tuning process, we conducted 2 independent training runs using the same hyperparameter configuration. These runs resulted in 2 models (FT1 and FT2), which were evaluated separately.

Embedding-Based Regression

To complement prompt- and fine-tuning-based approaches, we evaluated static embedding-based personality inference across 4 pretrained encoders of varying scales and training paradigms: all-mpnet-base-v2 (768-dimensional, built on masked and permuted pretraining for language understanding) [70] from the Sentence-Transformers library; E5-large-v2 (1,024-dimensional) [71], trained with contrastive

learning on large-scale text pairs and requiring a “*passage*” (prefix at inference time); and OpenAI text-embedding-3-small (1536-dimensional) and text-embedding-3-large (3072-dimensional), accessed via the OpenAI API [72]. This selection spans a broad range of model capacities and training paradigms, enabling systematic comparison among open-source encoders, contrastive-learning-based embeddings, and large-scale commercial models.

For all models, each participant’s interview transcript was encoded into a fixed-size dense vector at the transcript level. These embeddings, standardized to have 0 mean and unit variance, were passed to a multioutput ridge regression model ($\alpha=1.0$) [73] that jointly predicted all 5 Big Five trait scores. The dataset was split into 80% training and 20% testing partitions. We further conducted a sample-size sensitivity analysis by training the text-embedding-3-small ridge regression model on increasing proportions of the training pool (20%, 40%, 60%, 80%, and 100%) and evaluating performance on the same held-out test set. The 20% to 80% conditions were repeated 5 times with different random subsamples; the 100% condition used the full training pool.

Evaluation Metrics

We assessed model performance using 2 complementary metrics: Pearson correlation (r) and mean absolute error (MAE). Pearson correlation quantifies the consistency between model-predicted and self-reported trait scores, serving as an indicator of criterion-related validity. Correlations above $r=0.5$ are typically interpreted as stronger evidence of convergence with self-report scores, indicating that the model captures meaningful individual differences in personality. Values between 0.3 and 0.5 suggest moderate agreement with self-report scores, while correlations below 0.3 suggest weak predictive consistency. MAE provides a direct measure of prediction error on the original trait scale (1.0–5.0). Lower MAE values (eg, ≤ 0.5) reflect high precision, while values between 0.5 and 0.8 are common and may be acceptable depending on context. MAE values exceeding 1.0 typically signal poor calibration or unreliable estimates. Unlike correlation, MAE is sensitive to systematic bias and offers a more interpretable estimate of absolute prediction fidelity.

We also calculated sample-specific internal consistency for each-item trait scale and reported both Cronbach α and interitem correlations in Table A9 of [Multimedia Appendix 1](#).

Ethical Considerations

This study used deidentified semistructured interview transcripts and BFI-10 personality assessment data collected

from 5 prior Kent State University Institutional Review Board–approved studies under protocol numbers #17-056, #14-233, #13-134, #15-446, and #12-257 [60]. Participants provided informed consent in the original studies for the use of their data for research purposes. The present study involved secondary analysis of deidentified data and did not involve recontacting participants or collecting new identifiable information; therefore, additional informed consent was not required for this analysis.

To protect participant privacy and confidentiality, all textual and psychometric data were anonymized or deidentified before analysis and direct identifiers were not used in model training, evaluation, or reporting. Results are reported only in aggregate. Any interview excerpts included in the manuscript were shortened or edited to avoid revealing identifying details. No identifiable participant images, screenshots, or user-identifying information are included in the manuscript or supplementary materials. Participants in the original studies were financially compensated for their time at a rate of approximately US \$10 per hour. The total compensation varied across the parent studies because study activities and participation requirements differed; study-specific procedures are described in the corresponding published reports [74–78].

Given the sensitive nature of personality prediction, the models evaluated in this study are intended for research purposes only. They should not be used to identify individuals, make diagnostic judgments, or support high-stakes decisions without appropriate validation, ethical oversight, and human review.

Results

Prompt-Based Inference

Table 3 summarizes the performance of GPT-5 mini using zero-shot and CoT prompting. Overall, prompt-based methods exhibited weak and inconsistent performance across traits. The highest correlation was observed for *conscientiousness* in the zero-shot setting ($r=0.158$). Most other traits, including *neuroticism*, *openness*, and *extraversion*, showed correlations near 0 or were negative, indicating limited predictive validity. Prediction errors remained high across all traits. Only openness achieved a MAE below 0.7, whereas MAE values for conscientiousness and agreeableness exceeded 1.4, indicating poor calibration of predicted trait scores.

Table 3. Performance of prompt-based GPT-5 mini in predicting continuous Big Five personality scores from semistructured interviews in the pooled analytic sample (N=518).

Trait	Zero-shot (r^a)	CoT ^b (r)
Openness	–0.034	–0.010
Conscientiousness	0.158 ^c	0.155 ^c

Trait	Zero-shot (r^a)	CoT ^b (r)
Extraversion	-0.092 ^c	-0.092 ^c
Agreeableness	0.067	0.099 ^c
Neuroticism	-0.026	-0.034

^a r denotes the Pearson correlation coefficient.

^bCoT: chain-of-thought.

^cValues are statistically significant at $P<.05$. Detailed statistical results, including P values and CIs, are reported in Table A2 of [Multimedia Appendix 1](#).

LoRA-Based Adaptation

LoRA-based fine-tuning substantially improved personality prediction performance compared to prompt-based methods, with effects varying across traits and adaptation ranks.

As shown in [Table 4](#), performance gains were most pronounced for agreeableness, which showed statistically

significant correlations across all configurations ($r=0.299$ at rank 8, $r=0.329$ at rank 16, and $r=0.229$ at rank 32). Openness showed a different pattern, with performance improving as rank increased and reaching statistical significance only at rank 32 ($r=0.216$, $P<.05$).

Table 4. Performance of LLaMA-2-7B low-rank adaptation (LoRA) fine-tuning across adapter ranks for predicting continuous Big Five personality scores from semistructured interviews in the pooled analytic sample (N=518).

Trait	Rank 8 (r^a)	Rank 16 (r)	Rank 32 (r)
Openness	0.042	0.127	0.216 ^b
Conscientiousness	-0.057	0.032	-0.055
Extraversion	0.136	0.021	0.044
Agreeableness	0.299 ^b	0.329 ^b	0.229 ^b
Neuroticism	0.060	0.114	0.115

^a r denotes the Pearson correlation coefficient.

^bValues are statistically significant at $P<.05$. Full statistical results are reported in Table A3 of [Multimedia Appendix 1](#).

As shown in [Table 5](#), embedding-based LoRA models achieved stronger performance than other LoRA configurations. Conscientiousness achieved the highest correlation ($r=0.352$ at rank 8), which represents the best result within

this model family. Openness and neuroticism also showed relatively stable and statistically significant performance across ranks ($r\approx 0.23$ - 0.26).

Table 5. Performance of RoBERTa-base embedding-based low-rank adaptation (LoRA) fine-tuning across adapter ranks for predicting continuous Big Five personality scores from semistructured interviews in the pooled analytic sample (N=518).

Trait	Rank 8 (r^a)	Rank 16 (r)	Rank 32 (r)
Openness	0.258 ^b	0.251 ^b	0.255 ^b
Conscientiousness	0.352 ^b	0.143	0.232 ^b
Extraversion	0.147	0.143	0.197 ^b
Agreeableness	0.184	0.172	0.173
Neuroticism	0.246 ^b	0.232 ^b	0.239 ^b

^a r denotes the Pearson correlation coefficient.

^bValues are statistically significant at $P<.05$. Full statistical results are reported in Table A4 of [Multimedia Appendix 1](#).

In contrast, extraversion and agreeableness were generally weaker, with only extraversion, at rank 32, reaching statistical significance. Increasing the LoRA rank did not consistently improve performance, as some traits achieved their best results at lower ranks. Overall, these findings suggest that embedding-based models captured some personality-relevant signals more consistently, although performance varied across traits.

Supervised Fine-Tuning of GPT-4.1 mini

As shown in [Table 6](#), fine-tuning results were largely consistent for conscientiousness, agreeableness, and

neuroticism across the 2 runs. Conscientiousness showed the strongest performance ($r=0.686$ and 0.718), followed by agreeableness and neuroticism. Extraversion was less stable across runs ($r=0.389$ vs 0.172), while openness remained weak and nonsignificant ($r=-0.085$ and -0.138). These results suggest that fine-tuning more reliably captured certain personality traits, such as conscientiousness and agreeableness, while others remain more difficult to infer from text.

Table 6. Performance of GPT-4.1 mini supervised fine-tuning across 2 independent training runs for predicting continuous Big Five personality scores from semistructured interviews in the pooled analytic sample (N=518).

Trait	FT ^a 1 (<i>r</i> ^b)	FT2 (<i>r</i>)
Openness	−0.085	−0.138
Conscientiousness	0.686 ^c	0.718 ^c
Extraversion	0.389 ^c	0.172
Agreeableness	0.626 ^c	0.631 ^c
Neuroticism	0.422 ^c	0.403 ^c

^aFT: fine-tuning.
^b*r* denotes the Pearson correlation coefficient.
^cValues are statistically significant at *P*<.05. Full statistical results are reported in Table A5 of [Multimedia Appendix 1](#).

Embedding-Based Regression

As shown in [Table 7](#), embedding-based models achieved moderate, trait-dependent performance across pretrained encoders. Overall, text-embedding-3-large performed best,

yielding the highest correlations for conscientiousness (*r*=0.568), agreeableness (*r*=0.531), and neuroticism (*r*=0.249), all of which are statistically significant.

Table 7. Performance of embedding-based regression using pretrained text encoders for predicting continuous Big Five personality scores from semistructured interviews in the pooled analytic sample (N=518).

Trait	all-mpnet-base-v2	E5-large	text-embedding-3-small	text-embedding-3-large
Openness	−0.098	−0.036	0.189	0.060
Conscientiousness	0.174	0.386 ^a	0.476 ^a	0.568 ^a
Extraversion	0.302 ^a	0.132	0.102	0.250 ^a
Agreeableness	0.381 ^a	0.407 ^a	0.448 ^a	0.531 ^a
Neuroticism	0.040	0.096	0.119	0.249 ^a

^aValues are statistically significant at *P*<.05. Full statistical results are reported in Table A6 of [Multimedia Appendix 1](#).

Across traits, conscientiousness and agreeableness were the most consistently well-predicted, with significant correlations across multiple encoders. Extraversion showed modest-to-moderate performance, with significant results for all-mpnet-base-v2 (*r*=0.302) and text-embedding-3-large (*r*=0.250), but weaker performance for other models. In contrast, neuroticism showed limited performance, reaching statistical significance only with the largest embedding model. Openness remained difficult to predict across all encoders, with weak and nonsignificant correlations, consistent with patterns observed in other methods.

To examine whether predictive performance improved with larger training samples, we conducted a sample-size sensitivity analysis using the text-embedding-3-small, embedding-based regression model as a representative embedding baseline. Increasing proportions of the training pool were used for model training (20%, 40%, 60%, 80%, and 100%; corresponding to *n*=82, 165, 248, 331, and 414), while evaluation was performed on the same held-out test set. As shown in Table A8 of [Multimedia Appendix 1](#), increasing the training sample did not consistently improve predictive performance. Averaged across the 5 traits, Pearson correlations were 0.312, 0.326, 0.305, 0.300, and 0.267, respectively. At the trait level, conscientiousness and neuroticism generally decreased as the training sample increased, with their highest correlations observed at the 20% condition (*r*=0.581 and *r*=0.227, respectively), both higher than the corresponding full-training values. Agreeableness peaked at

the 40% condition (*r*=0.546) and then declined. In contrast, openness increased modestly across the training fractions, while extraversion increased through the 80% condition before declining at 100%; however, extraversion correlations remained below 0.20 across all training conditions. This pattern may partly reflect the use of a fixed ridge regularization parameter without systematic hyperparameter tuning or a separate validation set.

Discussion

Principal Findings

Across all modeling paradigms, personality prediction performance exhibits 2 consistent patterns: strong dependence on model adaptation strategy and substantial variation across personality traits. Consistent with our prior work on these data [24], prompt-based approaches (zero-shot and CoT prompting) yield weak and unstable performance. In contrast, supervised learning or embedding-based regression demonstrate clear improvements by enabling models to learn from labeled data. In addition, prediction performance varies across the Big Five traits, suggesting that different personality traits may benefit from different modeling approaches.

First, prompt-based inference methods without task-specific adaptation (zero-shot and CoT prompting) showed weak and unstable performance in directly predicting continuous Big Five personality scores from semistructured interview

transcripts. In contrast, Wright et al [30] reported moderate convergence between LLM-generated and self-reported personality trait scores based on spontaneous thought narratives and daily video diaries collected from 2 distinct samples. Compared with their unconstrained narrative data, our study evaluated personality prediction from semistructured interview transcripts, which may partly explain the lower zero-shot performance. In our benchmark, supervised approaches consistently outperformed prompt-based inference.

Second, performance varies systematically across traits. Conscientiousness and agreeableness are consistently the most predictable traits, achieving moderate-to-strong correlations across multiple methods, particularly under supervised fine-tuning ($r=0.718$ for conscientiousness and $r=0.631$ for agreeableness). Neuroticism shows moderate and relatively consistent performance, whereas extraversion is more variable across methods and runs. In contrast, openness shows the strongest and most consistent performance under embedding-based LoRA fine-tuning with RoBERTa-base ($r=0.251$ – 0.258), while LLaMA-2-7B LoRA achieves modest but statistically significant performance (best $r=0.216$). Prompt-based methods, supervised fine-tuning, and embedding-based regression remain comparatively weak, suggesting that prediction performance for openness is highly method-dependent.

Finally, increasing model complexity does not uniformly improve performance. While supervised fine-tuning provides the strongest overall results, parameter-efficient methods such as LoRA yield only modest gains, and embedding-based approaches offer a competitive trade-off between performance and efficiency. The sample-size sensitivity analysis did not show a clear performance gain within the current training range; however, larger cohorts and systematic variation in transcript length are still needed to determine the data conditions required for stable personality estimates.

These results are consistent with prior work showing that narrative interview language can be used to predict Big Five personality scores when language models are trained on validated psychometric measures [79]. They also support recent findings that free-form language may require psychologically grounded supervision to connect linguistic cues with questionnaire-based personality constructs [80].

The stronger performance of supervised fine-tuning, compared with prompting and most parameter-efficient configurations, indicates that mapping language representations to validated personality constructs requires direct optimization. At the same time, the competitive performance of embedding-based regression highlights that meaningful personality signals are already present in pretrained representations but require appropriate extraction mechanisms. In our implementation, transcripts are segmented into fixed-length chunks (512 tokens) and aggregated using mean pooling, which may smooth localized linguistic signals in longer conversational inputs.

The observed trait-specific differences further suggest that not all personality dimensions are equally expressed in

language. Traits such as conscientiousness and agreeableness appear to have stronger and more stable linguistic markers, whereas openness may depend on richer contextual or cognitive signals that are not fully captured by the available interview responses [81,82].

Limitations and Future Work

Although this study provides a systematic comparison of 4 modeling paradigms for personality inference from interview transcripts, several limitations constrain the generalizability of our findings.

First, although the dataset is based on real-world interview data, the sample size ($n=518$) remains modest relative to the parameter scale of large models, particularly decoder-based architectures such as LLaMA-2-7B. The sample-size sensitivity analysis did not show a clear performance gain within the current training range; however, future work should examine larger cohorts and systematically vary both participant sample size and transcript length to determine the data conditions needed for stable personality estimates. In addition, because the analytic sample pooled participants across occupational, clinical, student, and community cohorts, the findings may not generalize uniformly to any single population.

Second, several limitations also apply specifically to the embedding-based regression models. Variability in transcript length and the use of truncation, chunking, or mean pooling may lead to semantic information loss when long interview responses are compressed into fixed-size representations [83]. The ridge regression models were trained with a fixed regularization setting, without systematic hyperparameter tuning or a separate validation set for model selection; therefore, these results should be interpreted as baseline comparisons rather than fully optimized estimates of embedding-based model performance. Model performance was also evaluated using a single 80 to 20 train-test split rather than repeated k-fold cross-validation, which may make the reported estimates sensitive to the specific hold-out partition. Future work should use repeated cross-validation or external validation to better assess the robustness of these findings.

Third, the use of the short-form BFI-10 introduces inherent psychometric limitations. In the present sample, Cronbach α ranged from 0.429 for openness to 0.693 for conscientiousness, indicating nontrivial measurement noise in the self-reported criterion scores, which may attenuate observed model-criterion correlations and contribute to the apparent performance ceiling. It is also important to clarify that the criterion variables in this study are self-reported Big Five trait scores rather than personality in its full conceptual scope. Self-reports capture one form of personality-related information, partly reflecting individuals' explicit self-concepts and self-images [5]. Accordingly, LLM-derived estimates from interview language should be interpreted as predictions of self-reported trait scores, not as direct measurements of personality itself.

Fourth, our evaluation of prompting methods was limited to zero-shot and CoT strategies, which do not fully capture the broader space of prompt design strategies. Future studies should investigate few-shot prompting, trait-specific instructions, and prompting strategies that explicitly incorporate emotional or contextual signals, particularly for traits such as neuroticism.

Fifth, while we explored parameter-efficient adaptation using LoRA, our analysis was limited to variations in rank under a fixed configuration. Other factors, such as layer-wise adaptation, learning rate scheduling, and integration with retrieval-based methods (eg, RAG), may further improve performance, particularly for effectively salient traits such as neuroticism.

Finally, all models evaluated in this study were general-purpose pretrained models. Domain-adapted models trained on psychologically relevant or interview-based corpora may encode more informative representations for personality inference. Future work should compare language-based estimates with longer personality inventories, informant reports, behavioral indicators, and other criterion measures to evaluate whether they provide complementary information beyond self-report.

Conclusions

This study shows that personality inference from natural language is both feasible and fundamentally constrained,

revealing a pattern that is expressive but uneven and highly dependent on correspondence between language representations and psychometric constructs.

Contemporary LLMs encode meaningful personality-related signals but do not reliably translate these signals into calibrated trait estimates without explicit supervision. This pattern helps explain why models that appear linguistically fluent may nevertheless fail to produce stable or consistent personality predictions. Reliable personality inference therefore depends not only on model capability but also on how representations are aligned, adapted, and evaluated against validated psychological frameworks.

From a modeling perspective, personality prediction should be understood as a measurement problem rather than a purely generative task, in which performance is shaped by the interaction between data structure, model adaptation, and trait-specific linguistic expressivity. This reframing highlights that improving personality inference requires not only larger models but also a more principled integration of psychometric theory and representation learning.

Future work should extend this framework to more interactive and longitudinal settings, examining how personality signals evolve across extended conversations and diverse contexts. Additionally, incorporating richer personality measures and domain-adapted models may improve the reliability of trait-level inference.

Acknowledgments

During manuscript preparation, the generative AI (GenAI) tool ChatGPT (OpenAI) was used to assist with grammar checking, paragraph refinement, and clarity of wording. The authors reviewed, revised, and verified the accuracy of all AI-assisted content and take full responsibility for the accuracy, integrity, and final content of the manuscript. GenAI was not used to generate data, conduct analyses, or draw conclusions.

Funding

This study is a secondary analysis of data collected from multiple previously funded studies. Data collection for the Resilience and Adaptation to Injury study was supported by the National Institutes of Health (R01MH113622/R01MH113622-S1; principal investigator: KGC). The Psychological Adjustment to Multiple Sclerosis study was supported by a National Multiple Sclerosis Society Pilot Research Grant (principal investigator: KGC). The authors declared no financial support was received for the present study.

Data Availability

The data used in this study were collected under multiple Kent State University Institutional Review Board–approved protocols and contain sensitive human participant information. Therefore, the raw interview data are not publicly available. Qualified researchers may request access by contacting the corresponding author; requests will be reviewed in accordance with applicable institutional and ethical requirements. Documentation describing the pooled interview dataset is publicly available on Open Science Framework [60]. The analysis code is available online [61].

Authors' Contributions

Conceptualization: JZ, KGC, RJ

Formal analysis: JZ

Investigation: JZ, JM, XL

Methodology: JZ, JM, XL

Software: JZ

Supervision: KGC, RJ

Writing – original draft: JZ

Writing – review & editing: JZ, JM, XL, KGC, RJ

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompt templates, example model outputs, and supplementary statistical results.

[\[DOCX File \(Microsoft Word File\), 46 KB-Multimedia Appendix 1\]](#)

References

1. Wright AJ, Jackson JJ. Do changes in personality predict life outcomes? *J Pers Soc Psychol*. Dec 2023;125(6):1495-1518. [doi: [10.1037/pspp0000472](https://doi.org/10.1037/pspp0000472)] [Medline: [37384463](https://pubmed.ncbi.nlm.nih.gov/37384463/)]
2. Roberts BW, Kuncel NR, Shiner R, Caspi A, Goldberg LR. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect Psychol Sci*. Dec 2007;2(4):313-345. [doi: [10.1111/j.1745-6916.2007.00047.x](https://doi.org/10.1111/j.1745-6916.2007.00047.x)] [Medline: [26151971](https://pubmed.ncbi.nlm.nih.gov/26151971/)]
3. Demetriou A, Spanoudis G, Christou C, et al. Cognitive and personality predictors of school performance from preschool to secondary school: an overarching model. *Psychol Rev*. Mar 2023;130(2):480-512. [doi: [10.1037/rev0000399](https://doi.org/10.1037/rev0000399)] [Medline: [36315630](https://pubmed.ncbi.nlm.nih.gov/36315630/)]
4. Akrami N, Fernquist J, Isbister T, Kaati L, Pelzer B. Automatic extraction of personality from text: challenges and opportunities. Presented at: 2019 IEEE International Conference on Big Data (Big Data); Dec 9-12, 2019; Los Angeles, CA, USA. [doi: [10.1109/BigData47090.2019.9005467](https://doi.org/10.1109/BigData47090.2019.9005467)]
5. McDonald J. Measuring personality constructs: the advantages and disadvantages of self-reports, informant reports and behavioural assessments. University of Nottingham; 2008. URL: <https://www.nottingham.ac.uk/sociology/documents/enquire/volume-1-issue-1-dodorico-mcdonald.pdf> [Accessed 2026-10-03]
6. Spitzley LA, Wang X, Chen X, Burgoon JK, Dunbar NE, Ge S. Linguistic measures of personality in group discussions. *Front Psychol*. 2022;13:887616. [doi: [10.3389/fpsyg.2022.887616](https://doi.org/10.3389/fpsyg.2022.887616)] [Medline: [36186305](https://pubmed.ncbi.nlm.nih.gov/36186305/)]
7. Hampson SE. Personality processes: mechanisms by which personality traits “get outside the skin”. *Annu Rev Psychol*. 2011;63:315-339. [doi: [10.1146/annurev-psych-120710-100419](https://doi.org/10.1146/annurev-psych-120710-100419)] [Medline: [21740225](https://pubmed.ncbi.nlm.nih.gov/21740225/)]
8. Andresen PK, Schuurman NK, Hamaker EL. How to measure and model personality traits in everyday life: a qualitative analysis of 300 big five personality items. *J Res Pers*. Oct 2024;112:104528. [doi: [10.1016/j.jrp.2024.104528](https://doi.org/10.1016/j.jrp.2024.104528)]
9. Santos AMD, Moura FRT, Pinto LVL, et al. Words similarities on personalities: a language-based generalization approach for personality factors recognition. *IEEE Access*. 2023;11:29823-29836. [doi: [10.1109/ACCESS.2023.3261339](https://doi.org/10.1109/ACCESS.2023.3261339)]
10. Price LR. Psychometrics of drawmetrics: an expressive–semantic framework for personality assessment. *Behav Sci (Basel)*. Jan 17, 2026;16(1):135. [doi: [10.3390/bs16010135](https://doi.org/10.3390/bs16010135)] [Medline: [41595076](https://pubmed.ncbi.nlm.nih.gov/41595076/)]
11. Martínez-Huertas JÁ, Jorge-Botana G, Martínez-Mingo A, Moreno JD, Olmos R. Modeling personality language use with small semantic vector subspaces. *Pers Individ Dif*. Mar 2024;219:112514. [doi: [10.1016/j.paid.2023.112514](https://doi.org/10.1016/j.paid.2023.112514)]
12. Brickman J, Gupta M, Oltmanns JR. Large language models for psychological assessment: a comprehensive overview. *Adv Methods Pract Psychol Sci*. Jul 2025;8. [doi: [10.1177/25152459251343582](https://doi.org/10.1177/25152459251343582)]
13. Jiang G, Xu M, Zhu SC, Han W, Zhang C, Zhu Y. Evaluating and inducing personality in pre-trained language models. Presented at: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems; Dec 10-16, 2023; New Orleans, LA, USA. [doi: [10.5555/3666122.3666588](https://doi.org/10.5555/3666122.3666588)]
14. Naz A, Khan HU, Bukhari A, Alshemaimri B, Daud A, Ramzan M. Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges. *Artif Intell Rev*. 2025;58(8):239. [doi: [10.1007/s10462-025-11245-3](https://doi.org/10.1007/s10462-025-11245-3)]
15. Serapio-García G, Safdari M, Crepy C, et al. A psychometric framework for evaluating and shaping personality traits in large language models. *Nat Mach Intell*. 2025;7(12):1954-1968. [doi: [10.1038/s42256-025-01115-6](https://doi.org/10.1038/s42256-025-01115-6)] [Medline: [41438004](https://pubmed.ncbi.nlm.nih.gov/41438004/)]
16. Matšenas A, Lello A, Lees T, Peep H, Tamm KL. Can LLMs assess personality? Validating conversational AI for trait profiling. *arXiv*. Preprint posted online on Jan 23, 2026. [doi: [10.48550/arXiv.2602.15848](https://doi.org/10.48550/arXiv.2602.15848)]
17. Marengo D, Montag C, Settanni M. Inferring personality from social media activity using large language models: cross-model agreement, temporal stability, and convergent validity with self-reports. *J Pers*. Aug 2026;94(4):525-534. [doi: [10.1111/jopy.70019](https://doi.org/10.1111/jopy.70019)] [Medline: [40891650](https://pubmed.ncbi.nlm.nih.gov/40891650/)]
18. Stein JP, Messingschlager T, Gnambs T, Hutmacher F, Appel M. Attitudes towards AI: measurement and associations with personality. *Sci Rep*. Feb 5, 2024;14(1):2909. [doi: [10.1038/s41598-024-53335-2](https://doi.org/10.1038/s41598-024-53335-2)] [Medline: [38316898](https://pubmed.ncbi.nlm.nih.gov/38316898/)]
19. Piastra M, Catellani P. On the emergent capabilities of ChatGPT 4 to estimate personality traits. *Front Artif Intell*. 2025;8:1484260. [doi: [10.3389/frai.2025.1484260](https://doi.org/10.3389/frai.2025.1484260)] [Medline: [40017486](https://pubmed.ncbi.nlm.nih.gov/40017486/)]

20. Liu YF, Lu YL, He D, Zhang H. From five dimensions to many: large language models as precise and interpretable psychological profilers. Presented at: The Fourteenth International Conference on Learning Representations (ICLR 2026); Apr 23-27, 2026; Rio de Janeiro, Brazil. URL: https://proceedings.iclr.cc/paper_files/paper/2026/file/3d312e0c52a714b3c340507cd6bbf3e1-Paper-Conference.pdf [Accessed 2026-09-03]
21. Mehta Y, Fatehi S, Kazameini A, Stachl C, Cambria E, Eetemadi S. Bottom-up and top-down: predicting personality with psycholinguistic and language model features. Presented at: 2020 IEEE International Conference on Data Mining (ICDM); Nov 17-20, 2020; Sorrento, Italy. [doi: [10.1109/ICDM50108.2020.00146](https://doi.org/10.1109/ICDM50108.2020.00146)]
22. Dai Y, Jayaratne M, Jayatilleke B. Explainable personality prediction using answers to open-ended interview questions. *Front Psychol*. 2022;13:865841. [doi: [10.3389/fpsyg.2022.865841](https://doi.org/10.3389/fpsyg.2022.865841)] [Medline: [36467183](https://pubmed.ncbi.nlm.nih.gov/36467183/)]
23. Zhu J, Li X, Maharjan J, Coifman KG, Jin R. Benchmarking personality inference in large language models using real-world conversations. *J Psychiatry Brain Sci*. 2025;10(6):e250020. [doi: [10.20900/jpbs.20250020](https://doi.org/10.20900/jpbs.20250020)]
24. Zhu J, Jin R, Coifman KG. Can LLMs infer personality from real-world conversations? arXiv. Preprint posted online on Jul 18, 2025. [doi: [10.48550/arXiv.2507.14355](https://doi.org/10.48550/arXiv.2507.14355)]
25. Maharjan J, Jin R, Zhu J, Kenne D. Psychometric evaluation of large language model embeddings for personality trait prediction. *J Med Internet Res*. Jul 8, 2025;27:e75347. [doi: [10.2196/75347](https://doi.org/10.2196/75347)] [Medline: [40627556](https://pubmed.ncbi.nlm.nih.gov/40627556/)]
26. Yarkoni T. Personality in 100,000 words: a large-scale analysis of personality and word use among bloggers. *J Res Pers*. Jun 1, 2010;44(3):363-373. [doi: [10.1016/j.jrp.2010.04.001](https://doi.org/10.1016/j.jrp.2010.04.001)] [Medline: [20563301](https://pubmed.ncbi.nlm.nih.gov/20563301/)]
27. Dagha R, Dhaord R, Lakhani V, Hirlekar V. Personality prediction based on users' tweets. Presented at: 2022 2nd International Conference on Intelligent Technologies (CONIT); Jun 24-26, 2022; Hubli, India. [doi: [10.1109/CONIT55038.2022.9847670](https://doi.org/10.1109/CONIT55038.2022.9847670)]
28. Giorgi S, Nguyen KL, Eichstaedt JC, et al. Regional personality assessment through social media language. *J Pers*. Jun 2022;90(3):405-425. [doi: [10.1111/jopy.12674](https://doi.org/10.1111/jopy.12674)] [Medline: [34536229](https://pubmed.ncbi.nlm.nih.gov/34536229/)]
29. Peters H, Matz SC. Large language models can infer psychological dispositions of social media users. *PNAS Nexus*. Jun 2024;3(6):gae231. [doi: [10.1093/pnasnexus/pgae231](https://doi.org/10.1093/pnasnexus/pgae231)] [Medline: [38948324](https://pubmed.ncbi.nlm.nih.gov/38948324/)]
30. Wright AGC, Ringwald WR, Vize CE, et al. Assessing personality using zero-shot generative AI scoring of brief open-ended text. *Nat Hum Behav*. Mar 2026;10(3):541-555. [doi: [10.1038/s41562-025-02389-x](https://doi.org/10.1038/s41562-025-02389-x)] [Medline: [41617861](https://pubmed.ncbi.nlm.nih.gov/41617861/)]
31. John OP, Srivastava S. The Big Five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin LA, John OP, editors. *Handbook of Personality: Theory and Research*. 2nd ed. Guilford Press; 1999:102-138. URL: https://jenni.uchicago.edu/econ-psych-traits/John_Srivastava_1995_big5.pdf [Accessed 2026-10-03]
32. Kamalesh MD, Bharathi B. Personality prediction model for social media using machine learning technique. *Comput Electr Eng*. May 2022;100:107852. [doi: [10.1016/j.compeleceng.2022.107852](https://doi.org/10.1016/j.compeleceng.2022.107852)]
33. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. The development and psychometric properties of LIWC-22. University of Texas at Austin; 2022. URL: <https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf> [Accessed 2026-10-03]
34. Vishnubhotla K, Mohammad SM. Tweet emotion dynamics: emotion word usage in tweets from US and Canada. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association; 2022:4162-4176. URL: <https://aclanthology.org/2022.lrec-1.442/> [Accessed 2026-10-03]
35. Chen K, Zhang Z, Long J, Zhang H. Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Syst Appl*. Dec 2016;66(C):245-260. [doi: [10.1016/j.eswa.2016.09.009](https://doi.org/10.1016/j.eswa.2016.09.009)]
36. Guo A, Hirai R, Ohashi A, Chiba Y, Tsunomori Y, Higashinaka R. Personality prediction from task-oriented and open-domain human-machine dialogues. *Sci Rep*. Feb 16, 2024;14(1):3868. [doi: [10.1038/s41598-024-53989-y](https://doi.org/10.1038/s41598-024-53989-y)] [Medline: [38366048](https://pubmed.ncbi.nlm.nih.gov/38366048/)]
37. Huang C. Social media addiction and personality: a meta-analysis. *Asian J Social Psychol*. Dec 2022;25(4):747-761. [doi: [10.1111/ajsp.12531](https://doi.org/10.1111/ajsp.12531)]
38. Hinds J, Joinson AN. Digital data and personality: a systematic review and meta-analysis of human perception and computer prediction. *Psychol Bull*. Jun 2024;150(6):727-766. [doi: [10.1037/bul0000430](https://doi.org/10.1037/bul0000430)] [Medline: [38753373](https://pubmed.ncbi.nlm.nih.gov/38753373/)]
39. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: [10.48550/arXiv.1907.11692](https://doi.org/10.48550/arXiv.1907.11692)]
40. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on Oct 11, 2018. [doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)]
41. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. *Proc Natl Acad Sci U S A*. Apr 9, 2013;110(15):5802-5805. [doi: [10.1073/pnas.1218772110](https://doi.org/10.1073/pnas.1218772110)] [Medline: [23479631](https://pubmed.ncbi.nlm.nih.gov/23479631/)]
42. Gjrković M, Karan M, Vukojević I, Bošnjak M, Snajder J. PANDORA talks: personality and demographics on Reddit. In: *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics; 2021:138-152. [doi: [10.18653/v1/2021.socialnlp-1.12](https://doi.org/10.18653/v1/2021.socialnlp-1.12)]

43. Schwartz HA, Eichstaedt JC, Kern ML, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One*. 2013;8(9):e73791. [doi: [10.1371/journal.pone.0073791](https://doi.org/10.1371/journal.pone.0073791)] [Medline: [24086296](https://pubmed.ncbi.nlm.nih.gov/24086296/)]
44. Datta A, Chakraborty S, Mukherjee A. Personality detection and analysis using Twitter data. *arXiv*. Preprint posted online on Sep 11, 2023. [doi: [10.48550/arXiv.2309.05497](https://doi.org/10.48550/arXiv.2309.05497)]
45. Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J. Personalizing dialogue agents: I have a dog, do you have pets too? In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics; 2018:2204-2213. [doi: [10.18653/v1/P18-1205](https://doi.org/10.18653/v1/P18-1205)]
46. Lal YK, Nilsson A, Schwartz H. Systematic evaluation of GPT-3 for zero-shot personality estimation. In: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Association for Computational Linguistics; 2023:390-400. [doi: [10.18653/v1/2023.wassa-1.34](https://doi.org/10.18653/v1/2023.wassa-1.34)]
47. Ramirez A, Alsalihi M, Aggarwal K, Li C, Wu L, Walker M. Controlling personality style in dialogue with zero-shot prompt-based learning. *arXiv*. Preprint posted online on Feb 8, 2023. [doi: [10.48550/arXiv.2302.03848](https://doi.org/10.48550/arXiv.2302.03848)]
48. Rathje S, Mirea DM, Sucholutsky I, Marjeh R, Robertson CE, Van Bavel JJ. GPT is an effective tool for multilingual psychological text analysis. *Proc Natl Acad Sci U S A*. Aug 20, 2024;121(34):e2308950121. [doi: [10.1073/pnas.2308950121](https://doi.org/10.1073/pnas.2308950121)] [Medline: [39133853](https://pubmed.ncbi.nlm.nih.gov/39133853/)]
49. Yang T, Shi T, Wan F, et al. PsyCoT: psychological questionnaire as powerful chain-of-thought for personality detection. In: Bouamor H, Pino J, Bali K, editors. *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics; 2023:3305-3320. [doi: [10.18653/v1/2023.findings-emnlp.216](https://doi.org/10.18653/v1/2023.findings-emnlp.216)]
50. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. *arXiv*. Preprint posted online on Jun 17, 2021. [doi: [10.48550/arXiv.2106.09685](https://doi.org/10.48550/arXiv.2106.09685)]
51. Banerjee AM, Mukhopadhyaya K. Fine-tuning large language models for personality development. Presented at: 2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA); Nov 5-7, 2025; Coimbatore, India. [doi: [10.1109/ICECA66444.2025.11382960](https://doi.org/10.1109/ICECA66444.2025.11382960)]
52. Shen L, Long Y, Cai X, Chen G, Razzak I, Jameel S. Less but better: parameter-efficient fine-tuning of large language models for personality detection. Presented at: 2025 International Joint Conference on Neural Networks (IJCNN); Jun 30 to Jul 5, 2025; Rome, Italy. [doi: [10.1109/IJCNN64981.2025.11228339](https://doi.org/10.1109/IJCNN64981.2025.11228339)]
53. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv*. Preprint posted online on May 22, 2020. [doi: [10.48550/arXiv.2005.11401](https://doi.org/10.48550/arXiv.2005.11401)]
54. Zerhoubi S, Granitzer M. PersonaRAG: enhancing retrieval-augmented generation systems with user-centric agents. *arXiv*. Preprint posted online on Jul 12, 2024. [doi: [10.48550/arXiv.2407.09394](https://doi.org/10.48550/arXiv.2407.09394)]
55. Christian H, Suhartono D, Chowanda A, Zamli KZ. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *J Big Data*. Dec 2021;8(1):68. [doi: [10.1186/s40537-021-00459-1](https://doi.org/10.1186/s40537-021-00459-1)]
56. Ji Y, Wu W, Zheng H, Hu Y, Chen X, He L. Is ChatGPT a good personality recognizer? A preliminary study. *arXiv*. Preprint posted online on Jul 8, 2023. [doi: [10.48550/arXiv.2307.03952](https://doi.org/10.48550/arXiv.2307.03952)]
57. Sirasapalli JJ, Malla RM. A deep learning approach to text-based personality prediction using multiple data sources mapping. *Neural Comput Applic*. Oct 2023;35(28):20619-20630. [doi: [10.1007/s00521-023-08846-w](https://doi.org/10.1007/s00521-023-08846-w)]
58. Hu L, He H, Wang D, Zhao Z, Shao Y, Nie L. LLM vs small model? Large language model based text augmentation enhanced personality detection model. Presented at: AAAI'24/IAAI'24/EAAI'24: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence; Feb 20-27, 2024; Vancouver, British Columbia, Canada. [doi: [10.1609/aaai.v38i16.29782](https://doi.org/10.1609/aaai.v38i16.29782)]
59. Wang Y, Zhao J, Ones DS, He L, Xu X. Evaluating the ability of large language models to emulate personality. *Sci Rep*. 2025;15(1):519. [doi: [10.1038/s41598-024-84109-5](https://doi.org/10.1038/s41598-024-84109-5)]
60. Coifman K. Pooled experiential and diagnostic interviews of at-risk adults. Open Science Framework (OSF). 2026. URL: https://osf.io/357yk/overview?view_only=12566422d5394a6eb78b89e6e597fe01 [Accessed 2026-02-13]
61. LLM-personality-inference-interviews. GitHub. 2026. URL: <https://github.com/Sallyzhu/LLM-Personality-Inference-Interviews> [Accessed 2026-10-03]
62. Coifman KG, Bonanno GA. When distress does not become depression: emotion context sensitivity and adjustment to bereavement. *J Abnorm Psychol*. Aug 2010;119(3):479-490. [doi: [10.1037/a0020113](https://doi.org/10.1037/a0020113)] [Medline: [20677837](https://pubmed.ncbi.nlm.nih.gov/20677837/)]
63. Coifman KG, Flynn JJ, Pinto LA. When context matters: negative emotions predict psychological health and adjustment. *Motiv Emot*. Aug 2016;40(4):602-624. [doi: [10.1007/s11031-016-9553-y](https://doi.org/10.1007/s11031-016-9553-y)]
64. Harvey MM, Coifman KG, Ross G, Kleinert D, Giardina P. Contextually appropriate emotional word use predicts adaptive health behavior: emotion context sensitivity and treatment adherence. *J Health Psychol*. May 2016;21(5):579-589. [doi: [10.1177/1359105314532152](https://doi.org/10.1177/1359105314532152)] [Medline: [24801328](https://pubmed.ncbi.nlm.nih.gov/24801328/)]

65. GPT-5 mini model. OpenAI API. URL: <https://developers.openai.com/api/docs/models/gpt-5-mini> [Accessed 2026-05-10]
66. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Presented at: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; Dec 28-9, 2022; New Orleans, LA, USA. [doi: [10.5555/3600270.3601883](https://doi.org/10.5555/3600270.3601883)]
67. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Presented at: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; Dec 28-9, 2022; New Orleans, Louisiana, USA. [doi: [10.5555/3600270.3602070](https://doi.org/10.5555/3600270.3602070)]
68. Llama-2-7b-chat-hf. Hugging Face. 2024. URL: <https://huggingface.co/meta-llama/Llama-2-7b-chat-hf> [Accessed 2026-04-19]
69. GPT-4.1 mini model. OpenAI Developers. URL: <https://developers.openai.com/api/docs/models/gpt-4.1-mini> [Accessed 2026-04-09]
70. All-mpnet-base-v2. Hugging Face. 2024. URL: <https://huggingface.co/sentence-transformers/all-mpnet-base-v2> [Accessed 2026-04-16]
71. Wang L, Yang N, Huang X, et al. Text embeddings by weakly-supervised contrastive pre-training. arXiv. Preprint posted online on Dec 7, 2022. [doi: [10.48550/arXiv.2212.03533](https://doi.org/10.48550/arXiv.2212.03533)]
72. Text-embedding-3-small model. OpenAI Developers. URL: <https://developers.openai.com/api/docs/models/text-embedding-3-small> [Accessed 2026-04-16]
73. Ridge. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html [Accessed 2026-04-16]
74. Baugher BJ, Delahanty DL, Coifman KG. Recalled protective childhood experiences are associated with positive emotional expression in active-duty firefighters. *J Trauma Stress*. Feb 2024;37(1):187-194. [doi: [10.1002/jts.22992](https://doi.org/10.1002/jts.22992)] [Medline: [37955950](https://pubmed.ncbi.nlm.nih.gov/37955950/)]
75. Seah THS, Almahmoud S, Coifman KG. Feel to heal: negative emotion differentiation promotes medication adherence in multiple sclerosis. *Front Psychol*. 2022;12:687497. [doi: [10.3389/fpsyg.2021.687497](https://doi.org/10.3389/fpsyg.2021.687497)] [Medline: [35082708](https://pubmed.ncbi.nlm.nih.gov/35082708/)]
76. Mitchell BJ, Gawlik EA, Baugher BJ, et al. Were there losses in social support during the pandemic? Testing the impact of COVID-19 on psychological adjustment to trauma in United States adults. *Front Psychol*. 2022;13:1061621. [doi: [10.3389/fpsyg.2022.1061621](https://doi.org/10.3389/fpsyg.2022.1061621)] [Medline: [36619028](https://pubmed.ncbi.nlm.nih.gov/36619028/)]
77. Mitchell BJ, Aurora P, Coifman KG. Personality or pathology? Predictors of early substance use in first-year college students. *J Am Coll Health*. Jul 2023;71(5):1630-1637. [doi: [10.1080/07448481.2021.1947297](https://doi.org/10.1080/07448481.2021.1947297)] [Medline: [34243686](https://pubmed.ncbi.nlm.nih.gov/34243686/)]
78. Coifman KG, Aurora P. From social avoidance to substance use: working memory and negative affectivity predict maladaptive regulatory behaviors in daily life across diagnostic groups. *Cogn Ther Res*. Jun 2022;46(3):590-607. [doi: [10.1007/s10608-021-10284-8](https://doi.org/10.1007/s10608-021-10284-8)]
79. Oltmanns JR, Khandelwal R, Ma J, et al. Language-based AI modeling of personality traits and pathology from life narrative interviews. *J Psychopathol Clin Sci*. Jan 2026;135(1):122-135. [doi: [10.1037/abn0001047](https://doi.org/10.1037/abn0001047)] [Medline: [41021515](https://pubmed.ncbi.nlm.nih.gov/41021515/)]
80. Lyu Y, Zhang L. Ask, answer, and detect: role-playing LLMs for personality detection with question-conditioned mixture-of-experts. arXiv. Preprint posted online on Dec 9, 2025. [doi: [10.48550/arXiv.2512.08814](https://doi.org/10.48550/arXiv.2512.08814)]
81. Lukac M. Speech-based personality prediction using deep learning with acoustic and linguistic embeddings. *Sci Rep*. Dec 3, 2024;14(1):30149. [doi: [10.1038/s41598-024-81047-0](https://doi.org/10.1038/s41598-024-81047-0)] [Medline: [39627367](https://pubmed.ncbi.nlm.nih.gov/39627367/)]
82. Samadi MA, Nixon N. Personalities at play: probing alignment in AI teammates. arXiv. Preprint posted online on Feb 28, 2026. [doi: [10.48550/arXiv.2603.00429](https://doi.org/10.48550/arXiv.2603.00429)]
83. Chhoun S, Po P, Ros S, Cho WS, Khoeurn S. Evaluation of chunking strategies for effective text embedding in low-resource language on agricultural documents. arXiv. Preprint posted online on May 21, 2026. [doi: [10.48550/arXiv.2605.22203](https://doi.org/10.48550/arXiv.2605.22203)]

Abbreviations

BERT: bidirectional encoder representations from transformers
BFI-10: 10-item Big Five Inventory
CoT: chain-of-thought
LLaMA: Large Language Model Meta AI
LLM: large language model
LoRA: low-rank adaptation
MAE: mean absolute error
OSF: Open Science Framework
RAG: retrieval-augmented generation

RoBERTa: Robustly Optimized BERT Approach

Edited by Andrew Coristine; peer-reviewed by Conal Monaghan, Weilin Xu; submitted 13.Nov.2025; final revised version received 23.Aug.2026; accepted 25.Aug.2026; published 16.Oct.2026

Please cite as:

Zhu J, Maharjan J, Li X, Coifman KG, Jin R

Personality Inference From Semistructured Interviews Using Large Language Models: Comparative Evaluation Study

J Med Internet Res 2026;28:e87129

URL: <https://www.jmir.org/2026/1/e87129>

doi: [10.2196/87129](https://doi.org/10.2196/87129)

© Jianfeng Zhu, Julina Maharjan, Xinyu Li, Karin G Coifman, Ruoming Jin. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 16.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.