

Do Large Language Models (LLMs) Really Understand Personality? A Test of Embeddings vs. Zero-Shot

Julina Maharjan, Ruoming Jin, Jianfeng Zhu, Deric Kenne

Submitted to: Journal of Medical Internet Research
on: April 01, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	4
Supplementary Files.....	25

Preprint
JMIR Publications

Do Large Language Models (LLMs) Really Understand Personality? A Test of Embeddings vs. Zero-Shot

Julina Maharjan¹ PhD; Ruoming Jin¹ Prof Dr; Jianfeng Zhu¹ PhD; Deric Kenne² Prof Dr

¹ Department of Computer Science, Kent State University Kent US

² Department of Public Health Kent State University Kent US

Corresponding Author:

Julina Maharjan PhD

Department of Computer Science, Kent State University
800 E. Summit St.
Kent
US

Abstract

Recent advancements in Large Language Models (LLMs) have sparked interdisciplinary interest in their ability to assess psychological constructs, particularly Personality. While prior machine learning research has focused on evaluating LLMs' capability to infer personality traits, often via zero-shot or few-shot learning, few studies have systematically examined the applicability of LLM embeddings for Personality Prediction within a robust psychometric validity framework or explored their correlation with psychological and linguistic features. Addressing this gap, we investigate performance of LLM embeddings on a well-labeled PANDORA dataset (Big Five Personality traits from Reddit). Our key contributions are: (1) demonstrating that LLM embeddings significantly outperform zero-shot approaches in personality prediction; (2) validating the reliability of embeddings using psychometric evaluation and analyzing their correlations with linguistic features (e.g., LIWC, emotional markers); (3) contrasting embedding performance against advanced feature engineering methods (joint and independent use of psycholinguistic features); and (4) revealing that model size influences efficacy, OpenAI outperforming RoBERTa. Our findings underscore the potential of LLM embeddings as a scalable and interpretable tool for computational personality assessment.

(JMIR Preprints 01/04/2025:75347)

DOI: <https://doi.org/10.2196/preprints.75347>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR journal, I will be able to make the full manuscript visible to all users.

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Do Large Language Models (LLMs) Really Understand Personality? A Test of Embeddings vs. Zero-Shot

Abstract

Recent advancements in Large Language Models (LLMs) have sparked interdisciplinary interest in their ability to assess psychological constructs, particularly Personality. While prior machine learning research has focused on evaluating LLMs' capability to infer personality traits, often via zero-shot or few-shot learning, few studies have systematically examined the applicability of LLM embeddings for Personality Prediction within a robust psychometric validity framework or explored their correlation with psychological and linguistic features. Addressing this gap, we investigate performance of LLM embeddings on a well-labeled PANDORA dataset (Big Five Personality traits from Reddit). Our key contributions are: (1) demonstrating that LLM embeddings significantly outperform zero-shot approaches in personality prediction; (2) validating the reliability of embeddings using psychometric evaluation and analyzing their correlations with linguistic features (e.g., LIWC, emotional markers); (3) contrasting embedding performance against advanced feature engineering methods (joint and independent use of psycholinguistic features); and (4) revealing that model size influences efficacy, OpenAI outperforming RoBERTa. Our findings underscore the potential of LLM embeddings as a scalable and interpretable tool for computational personality assessment.

Keywords: personality; Big5; DeepLearning, LLMs; embeddings; social media; reddit

Introduction

Personality, a foundational construct in psychology, systematically influences human behavior, cognition, and emotion (Murray, H. A. (1936)) [1]. Its assessment has become indispensable across both academic and applied domains. Contemporary applications range from clinical diagnostics to AI-driven applications such as job candidate evaluation, mental health screening, and personalized recommendation system (Goodstein, L. D. 1999) [2]. This widespread utility has driven innovation in assessment methodologies, particularly as traditional self-report measures face well-documented limitations regarding scalability and objectivity.

The field has increasingly turned to language-based assessment approaches, grounded in the robust theoretical framework of the Big Five personality model (Costa & McCrae, 1992) [3]. This empirically validated taxonomy organizes personality into five core dimensions, Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. These dimensions provide a comprehensive yet parsimonious framework for understanding individual differences, making them particularly suitable for computational approaches.

Openness: Represents an individual's level of curiosity, imagination, and willingness to try new things.

Conscientiousness: Reflects how organized, responsible, and detail-oriented an individual is.

Extraversion: Indicates how outgoing, sociable, and energetic an individual is.

Agreeableness: Measures how cooperative, trusting, and kind an individual is.

Neuroticism: Represents the level of emotional instability, anxiety, and mood swings an individual experiences.

Social media platforms have emerged as particularly valuable sources for personality assessment, offering vast quantities of natural language data that capture authentic behavioral expressions. Unlike controlled laboratory settings or structured surveys, these digital traces provide ecologically valid indicators of personality in real-world contexts. However, the unstructured nature of such data presents significant analytical challenges, including linguistic noise, contextual variability, and potential cultural biases, challenges that demand sophisticated computational solutions.

Recent advances in AI, particularly in large language models (LLMs), have introduced transformative capabilities for personality assessment in natural language data. These models can generate dense, context-aware embeddings that potentially capture deeper psychosocial patterns than traditional psycholinguistic features. Their ability to process natural language at scale while maintaining sensitivity to contextual nuances suggests they may overcome many limitations of previous approaches. In addition to this, the emergence of conversational AI agents like ChatGPT [4] and DeepseekAI [5] has revolutionized this domain by enabling zero-shot personality inference, the ability to assess traits without task-specific training. These models can generate preliminary personality profiles through direct prompting, offering immediate applicability in real-world settings. However, while zero-shot approaches demonstrate promising scalability, their psychometric properties and depth of analysis remain uncertain compared to more systematic methods. This uncertainty motivates our investigation of whether purpose-trained embedding models can surpass zero-shot capabilities while maintaining practical utility. Further, the field currently lacks comprehensive validation of these methods against established psychometric standards, particularly regarding reliability, construct validity, and practical utility across different assessment contexts.

Our study addresses these critical gaps through systematic evaluation of LLM embeddings using the PANDORA dataset, a carefully curated collection of Reddit posts annotated with Big Five personality traits. Our investigation focuses on four critical research questions: first, we examine the predictive efficacy of LLM embeddings trained on custom deep learning models and compare it with zero-shot approach; second, we assess LLM embeddings using established validation frameworks and explore their interpretability through correlations with psychological properties like LIWC and emotion features; and third, we examine the effectiveness of traditional psycholinguistic features and advanced feature engineering along with LLM embeddings, and lastly, we examine performance difference across leading LLM architectures (e.g. OpenAI, BERT, RoBERTa).

By grounding this work in the comprehensive Big Five model, we ensure theoretical consistency while advancing computational methods for personality assessment, ultimately contributing to both psychological science and applied AI development.

Literature Review

The rapid evolution of Large Language Models (LLMs) has fundamentally transformed personality prediction research, offering new capabilities through both conversational agents [15] and embedding-based approaches [7-11]. Recent work has established that LLMs not only exhibit emergent personality-like traits but can also emulate specific personality dimensions through targeted prompting techniques [16-18]. This capability stems from their advanced semantic understanding and contextual awareness, as shown Jiang et al.'s (2023) examination of personality induction in pre-trained models. However, the field lacks consensus on appropriate validation frameworks for these emerging methods. While studies such as Serapio-García et al. (2023) [19] and [20] have proposed methodologies for quantifying personality in LLMs, their focus has primarily been on direct model outputs rather than the embedding representations that underlie these predictions.

Psychometric Foundations for Computational Assessment

Personality assessment in psychological science relies on rigorous validation of reliability (internal consistency) and validity (convergent/discriminant) [John & Soto, 2021] [21]. Although recent AI studies have adopted these principles [16-20], most examine only LLMs' direct responses to structured inventories rather than their embeddings' psychometric properties. Critical gaps remain in assessing whether embeddings: (1) maintain measurement invariance across demographics [Boyd & Pennebaker, 2017] [26], (2) align with theoretical constructs beyond surface lexical patterns, and (3) demonstrate consistent reliability in real-world applications. Our work addresses these gaps through comprehensive validation against established psychometric frameworks while correlating embeddings with psychological lexicons (LIWC, emotion markers) for theoretical grounding.

The Evolution of Text-Based Personality Prediction

The computational assessment of personality has progressed through three methodological generations. Early work (2000–2015) established the validity of handcrafted linguistic features, most notably the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2001) [22], which demonstrated 0.38–0.42 correlations with Big Five traits in Facebook data (Schwartz et al., 2013) [27]. Subsequent feature sets like Mairesse traits (2007) [23] and NRC Emotion Lexicon (Mohammad & Turney, 2013) [24] enabled more granular analyses, with Park et al. (2015) [28] achieving 61% accuracy in Openness prediction using LIWC+NRC hybrid features. However, these approaches faced well-documented limitations in handling contextual language use (Tausczik & Pennebaker, 2010) [25], particularly for constructs like Neuroticism where sarcasm and irony distorted predictions (Hovy, 2015) [30].

The transformer revolution (2018–present) initially focused on BERT variants, with RoBERTa-base improving LIWC's performance by 19.3% F1 on the PANDORA dataset (Gjurković et al., 2020) [6]. Despite these advances, critical limitations persist in the literature. Serapio-García et al. [19] identified three major shortcomings: first, the predominant use of generic pretrained models (e.g., vanilla BERT-base) without domain-specific adaptation for personality assessment; second, insufficient controlled comparisons against established psycholinguistic features (Tausczik & Pennebaker, 2010; Schwartz et al., 2013); [25, 27]; and third, a concerning absence of measurement invariance testing across demographic groups, despite well-documented lexical biases in language-based assessments [26]. While recent hybrid approaches combining LIWC and BERT features have shown empirical promise [29], these methods often lack robust theoretical foundations explaining their feature integration strategies, potentially limiting their interpretability and generalizability. This gap underscores the need for more principled approaches to combining traditional and neural features in computational personality assessment.

Benchmarking Model Architectures

Current research on personality prediction architectures remains disproportionately focused on BERT-family models [31, 32], despite emerging evidence of their limitations. While BERT-large achieves strong performance (F1=0.682) on standard benchmarks, this narrow focus overlooks critical gaps: the efficiency-accuracy tradeoffs of larger models like RoBERTa (355M parameters) versus distilled versions, the untested potential of conversational LLMs like GPT-3.5 whose embeddings may capture richer psychosocial patterns, and the lack of comprehensive evaluations comparing zero-shot versus fine-tuned approaches, assessing cross-model reliability, or examining demographic fairness. These omissions persist despite known biases in language-based personality assessment and the growing need for deployable, equitable systems. Our work addresses these gaps through systematic comparison of architectures spanning different sizes and training paradigms

Our work addresses these gaps through systematic comparison of OpenAI's GPT-3.5 (175B),

representing conversational LLMs with zero-shot inference; and RoBERTa-large (355M), the current psycholinguistic standard. We extend Jiang et al.'s (2023) [16] prompt-based analysis by evaluating embeddings, while incorporating Levine et al.'s (2022) [31] cost-aware evaluation framework - the first such application in personality computing.

Research Questions

RQ1. Does LLMs' embedding provide better performance?

RQ2. How do LLM embeddings correlate with psychological constructs to enable personality prediction?

RQ3. Do traditional features (like LIWC) along with feature engineering improve prediction?

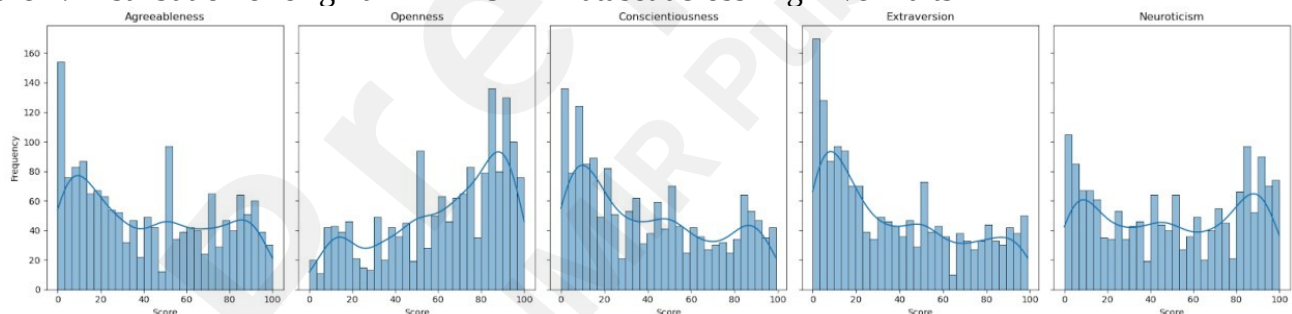
RQ4: How different LLM embedding works, which one is better?

Methodology

Data Collection and Preprocessing

We used PANDORA dataset which is a reddit dataset that comprises 17M posts from 10K users. The dataset contains 3 different kinds of personality models' data: namely MBTI, Big 5 and Enneagram. As our research is based on only Big 5 personality traits, we filtered out non-Big 5 data that resulted in 1,568 users and 3M Big 5 Personality posts. We further filtered out non-English posts that resulted in 2,830,311 posts. Many posts were short. Hence, we filtered out all the posts that were shorter than 30 words, which left us 1,008,627 posts. Additionally, we also filtered out duplicate posts that resulted in 935,102 posts with an average of 90-word counts. The distribution of the scores in each Big 5 Traits is shown in Figure 1.

Figure 1: Distribution of original PANDORA Dataset across Big Five Traits



The Big Five personality trait labels in the data were labeled with 3 different scoring methods: scores, percentiles and description, by the author ranging from 0 to 100. Exactly, there were 520065, 268189, and 145848 posts with scores, percentiles and description methods respectively. The average scores were found to be 43.9 for Agreeableness, 63.6 for Openness, 39.5 for Conscientiousness, 40.4 for Extraversion and 47.9 for Neuroticism.

Contextual Embedding from Large Language Models (LLMs)

Embeddings are dense and low dimensional vector representations of data (words or sentences or documents) that capture semantic meaning and relationships. When using Large Language Models like GPT, BERT, or RoBERTa, embeddings are generated as a byproduct of the model's internal processing. These embeddings can be used as features for downstream tasks like classification in our case. Unlike traditional vector embedding models like word2vec, embeddings from LLMs encode contextual meaning of words; meaning the same word can have different embeddings depending on its context. LLMs generate embeddings through their internal architecture, which is based on

transformers. The key steps in generating embeddings are tokenization, contextual encoding, and optional pooling.

Firstly, the input is split into tokens and mapped to an initial embedding vector from the model's embedding layer, known as tokenization. Then, the tokens' embeddings are processed through the LLM's transformer architecture, self-attention and feed-forward networks. Self-attention is a key technique where each token (Query) attends all neighboring tokens (Key) to generate actual result token (Value). This step is followed by FFNNs in multiple rounds, where a pooling is applied on the final layer. In our experiment, we pooled [CLS] token from the final layer of the model (RoBERTa, BERT, XLNet and TweetBERT) to represent the corresponding contextual sentence embedding.

While the process of generating embeddings from opensource pretrained Transformers based required specifying above 3 key steps, recently popular LLMs like OpenAI and DeepseekAI require to have paid APIs to generate embeddings. In our study, we used an opensource hub, Huggingface Transformers library to generate embeddings from RoBERTa as roberta-base, BERT as bert-base-uncased, XLNet and TweetBERT models, and used custom API for OpenAI and DeepseekAI. The embedding size of Huggingface transformers is 768. For the OpenAI embeddings, we subscribed to paid version and used the "text-embedding-small-

3" model for generating embeddings. The default dimension of OpenAI embedding is 1568. Table 1. Embedding Dimension across different LLM

	RoBERTa	BERT	OpenAI
Dimension	768	768	1568

RQ1. Does LLMs' embedding provide better performance?

Our primary work begins with experimenting the variant of LLMs' embedding (RoBERTa and OpenAI) to perform Big 5 Personality Traits prediction on our custom model (BiLSTM with Attention). To compare with SOTA [Killian Jr] [8], we have approached the regression task with Mean Square Error (MSE) as an evaluation metric. In addition to training the model using LLM embedding as features, we also experiment the zero-shot inference on the same test data. In the following sections, we explain the architecture of a custom classifier along with training settings, and prompt used for zero shot inference. Finally, we present the MSE results from custom classifier with RoBERTa and OpenAI embedding, and zero-shot inference.

Method

Training on a custom classifier

We extracted embeddings from Large Language Models, RoBERTa, OpenAI and DeepseekAI, as explained in Methodology section, and used them to train a classifier module; BiLSTM with Attention Model.

BiLSTM with Attention

The classifier module is built on BiLSTM (Bidirectional Long Short-Term Memory) Neural Network with Attention Mechanism, which is built on 512 hidden units. The two independent LSTMs; left and right captures the forward and backward information addressing the limitations of unidirectional models. The forward information h can be calculated as:

$$f_t = \text{sigmoid} (W_f [h_{(t-1)}, x_t] + b_f)$$

$$i_t = \text{sigmoid} (W_i [h_{(t-1)}, x_t] + b_i)$$

$$C_t' = \tanh (W_c [h_{(t-1)}, x_t] + b_c)$$

$$C_t = f_t C_{t-1} + i_t C_t'$$

$$o_t = \text{sigmoid} (W_o [h_{t-1}, x_t] + b_o)$$

$$h^f = o_t \tanh (C_t)$$

At each time t , x_t represents input, f_t , i_t , C_t , o_t represent the forget gate, input gate, cell state, and output gate respectively. W and b represent the weight and bias at its corresponding gates. Conversely, the backward h^b is also calculated. The overall output at t is provided by:

$$h_t = \text{sigmoid} (W_h [h^f, h^b] + b)$$

However, standard BiLSTM often struggle with long-range dependencies, leading us to incorporate an Attention mechanism that dynamically assigns importance to relevant textual features. Hence, we incorporated simple dot product attention layer to capture the weightage information from neighboring vectors; (context, h_c) for each output vector; h_t .

$$\text{Attention}(h_t, h_c) = \frac{\exp(\text{score}(h_t, h_c))}{\sum_{c'=1}^n \exp(\text{score}(h_t, h_{c'}))}$$

The attention layer computes weights over the hidden states of the BiLSTM, enabling the model to focus on crucial aspects of the input while mitigating information dilution. Finally, the output dense layer generates the score for each sequence vectors.

Training

We used *Cuda cuda_12.1.r12.1/compiler.32415258_0* version to train the dataset in GPU machine. For our model development, we used Sklearn, Pytorch and Huggingface libraries. The code base for the entire experimental design can be found in GitHub [34]. We begin with the training by splitting the dataset into 3 splits, 80:10:10 as Training set, Validation set and Testing set. Further for the advanced optimization, we leverage a Grid Search method on the validation dataset to search the best hyper-parameters for our classifiers. The best parameters for each classifier are presented in the Multimedia Appendix Table.

Zero-Shot Inference

For the same set of test dataset, we also present zero-shot inference scores. We used the GPTs' latest model (GPT4o) with a specific prompt design to infer the personality score for each input sequence. We specify the agent as psychologist expert and give the thorough features regarding Big 5 Personality traits, followed by the instruction for rating. Each sequence is rated on a scale of 5 (Very high) to 1 (Very Low). The detail structure of the prompt is presented in the table below:

“

You are an expert psychologist specializing in personality analysis.
Based on the Big Five Personality Traits model, you will evaluate an individual's responses to five questions.

Each response reflects different dimensions of personality traits.

For each of the Big Five traits, consider the following facets.

Conscientiousness: order, dutifulness, achievement striving, self-discipline, deliberation.

Agreeableness: trust, straightforwardness, altruism, compliance, modesty, tendermindedness.

Neuroticism: anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability.

Openness: fantasy, aesthetics, values.

Extraversion: warmth, gregariousness, assertiveness, excitement-seeking.

Instructions:

Read the individual's responses to five questions carefully.

For each personality trait, assign a score between 1 (Very Low) and 5 (Very High) based on the themes, tone, and content of the responses.

Each score can be rounded to the nearest tenth, like 3.5.

Scoring Guide:

1: Very Low - The response shows little to no alignment with the trait's facets. 2: Low - The response shows weak alignment with the trait's facets.

3: Moderate - The response shows some alignment but not strongly.

4: High - The response strongly aligns with the trait's facets.

5: Very High - The response shows exceptional alignment with the trait's facets. {questions_text}

Conversation: {conversation}

Task:

Analyze the given text for Big Five personality traits and provide the scores in the following format without any explanation or extra words:

"Conscientiousness": score, "Agreeableness": score, "Neuroticism": score, "Openness": score, "Extraversion": score

“

Evaluation Metric: Mean Square Error

Mean Squared Error (MSE) measures the average squared difference between the predicted values and the actual values. It is commonly used metric in regression tasks to evaluate the performance of a model. The smaller the MSE, the closer the model's predictions are. The mathematical notation for MSE is as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Where n is the number of data points, and y_i and $\hat{y}_i$ represents is the actual and predicted value respectively, for the i^{th} observation.

Result

Table 2. Performance measured by MSE across different LLM embeddings and zero-shot

	OPN	CON	EXT	AGR	NEU	Average
BERT (Killian Jr [8])	424.9	550.32	513.22	585.18	581.91	531.11

RoBERTa	496.12	641.38	610.37	702.071	641.64	618.44
OpenAI	417.87	546.43	510.37	599.50	560.33	526.90
zero-shot (GPT4.o)	668.81	767.038	848.35	919.47	624.57	765.647

Our experiments demonstrate that OpenAI embeddings trained on a BiLSTM model achieve superior performance (MSE = 526.90) compared to both other embedding approaches and zero-shot inference (MSE = 765.64). This substantial performance gap (45% lower error) not only confirms the advantage of custom-trained embeddings over direct zero-shot application but also suggests potential misinterpretation risks when relying solely on untuned LLM outputs for personality assessment.

RQ2. How do LLM Embeddings correlate with psychological constructs to enable personality prediction?

Reliability, Convergent and Divergent Validities on LLMs

In traditional questionnaire based Psychometric Tests, researchers often formulate the scales, a set of items to measure a latent construct as a prior task. For the personality measurement, psychologist formulates Big Five Inventory (BFI) consisting of subsequent items to be measured and undergoes validation check using psychometric test framework. The validation mainly ought to measure 1) the construct validity to test if scales reflect the underlying construct, 2) Reliability (Internal Consistency) to confirm if the scales are consistent over the same constructs, 3) Convergent Validity to confirm if the construct correlates with purported indicators (i.e., convergent tests) of the same or similar psychological construct, and 4) Divergent Validity to test the uncorrelation with the scores on theoretically unrelated tests.

Similar to item-based measurement, in our study, we aim to measure latent LLMs' embeddings to validate and check reliability. We performed validations on the predictions resulted from RQ1.

Cronbach alpha for Internal Consistency/ Reliability Test

Cronbach's alpha is a way of assessing reliability by comparing the amount of shared variance, or covariance, among the items making up an instrument to the amount of overall variance. Here, we evaluate the Cronbach alphas on the LLMs embeddings across all five-personality trait and observed the score above 0.6 in all personality traits except Conscientiousness.

Table 3. Cronbach score for LLMs embedding across Big Five Traits

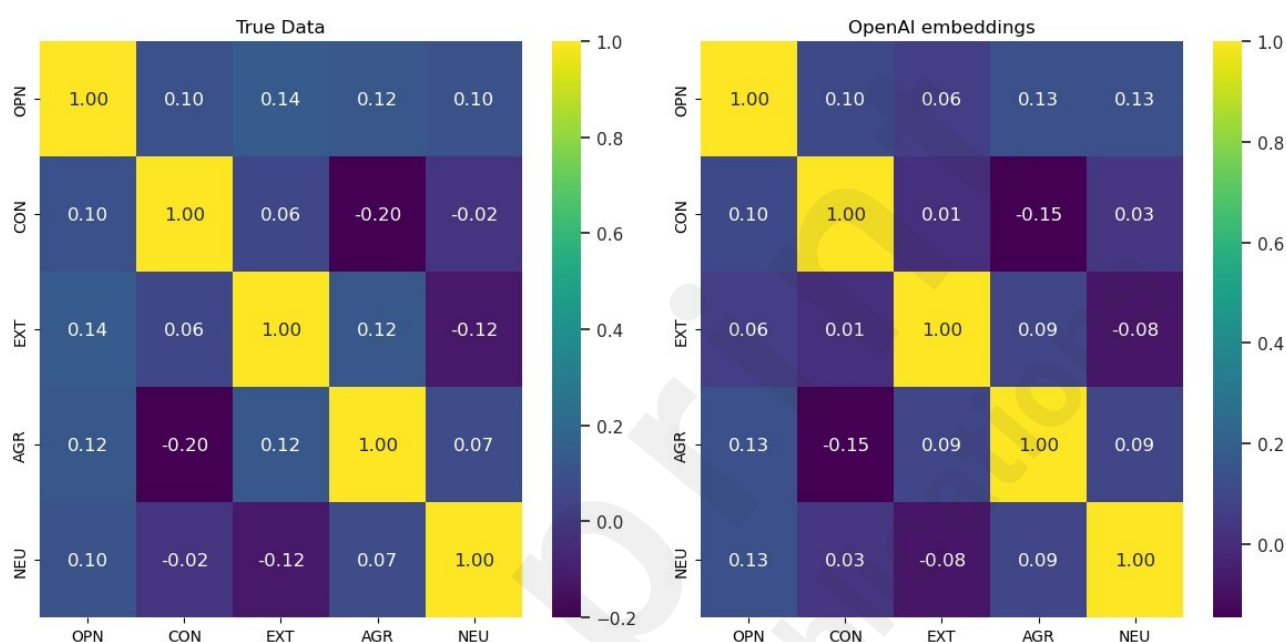
Openness	(0.6637411384440958, array([0.648, 0.679]))
Conscientiousness	(0.5742894806444646, array([0.547, 0.601]))
Extraversion	(0.6187154356717698, array([0.597, 0.64]))
Agreeableness	(0.6476180569410122, array([0.625, 0.669]))
Neuroticism	(0.6376958477093122, array([0.594, 0.679]))

Correlation Score for Convergent and Divergent Validity Test

Additionally, we calculated the correlation scores among Big 5 personality traits in two different ways: the actual score correlation and the embedding correlations across personality traits as shown

in the Figure [2] in the left and right respectively. Ideally, Openness and Agreeableness, and Conscientiousness and Extraversion are considered correlated, while Neuroticism is often unrelated with all other traits [33] by Oz, Huseyin. Our correlation results in both settings demonstrate a similar result. Also, the correlation score in both results have same direction except for the conscientiousness and neuroticism, where -0.02 was observed in true data while 0.003 in LLMs embeddings score.

Figure 2. Heatmap between Big Five Traits (a) Personality score in the true data (b) OpenAI embeddings



Our psychometric validation of LLM embeddings revealed strong alignment with traditional personality measurement principles. The embeddings demonstrated acceptable internal consistency (Cronbach's $\alpha > 0.6$) for four of the five traits, with only Conscientiousness falling slightly below this threshold. Correlation analyses showed remarkable convergence between embedding-derived traits and theoretical expectations: Openness-Agreeableness and Conscientiousness-Extraversion showed positive relationships, while Neuroticism remained largely independent of other traits, mirroring patterns in ground-truth data (Figure 2). The sole deviation appeared in the Conscientiousness-Neuroticism relationship, where embeddings showed near-zero correlation (0.003) versus the slight negative association (-0.02) in original scores, suggesting minimal but measurable divergence in how these traits relate in the embedding space.

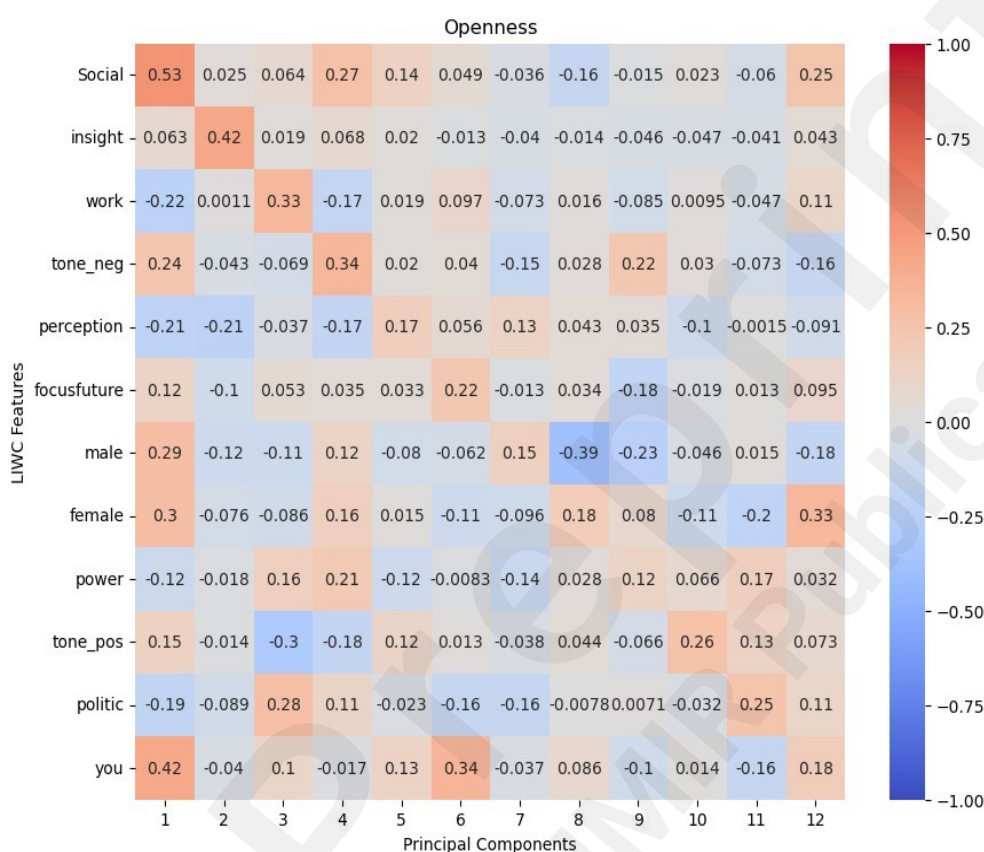
Correlation between LIWC Features and reduced LLM (OpenAI) embedding

Linguistic Inquiry and Word Count (LIWC-22), by Pennebaker, is the powerful linguistic analysis tool that allows researchers to understand the reflection of language on psychological and social dimensions. The tool helps in analyzing the texts by mapping the words in the text to pre-validated dictionaries of words from psychological categories like emotion, cognition, social process, and social references. In the personality trait assessment, the tool has been a primary choice among many researchers given its powerful psychological dictionary. In our work, we utilized this software to generate the static linguistic features (total of 119 features) on the studied data.

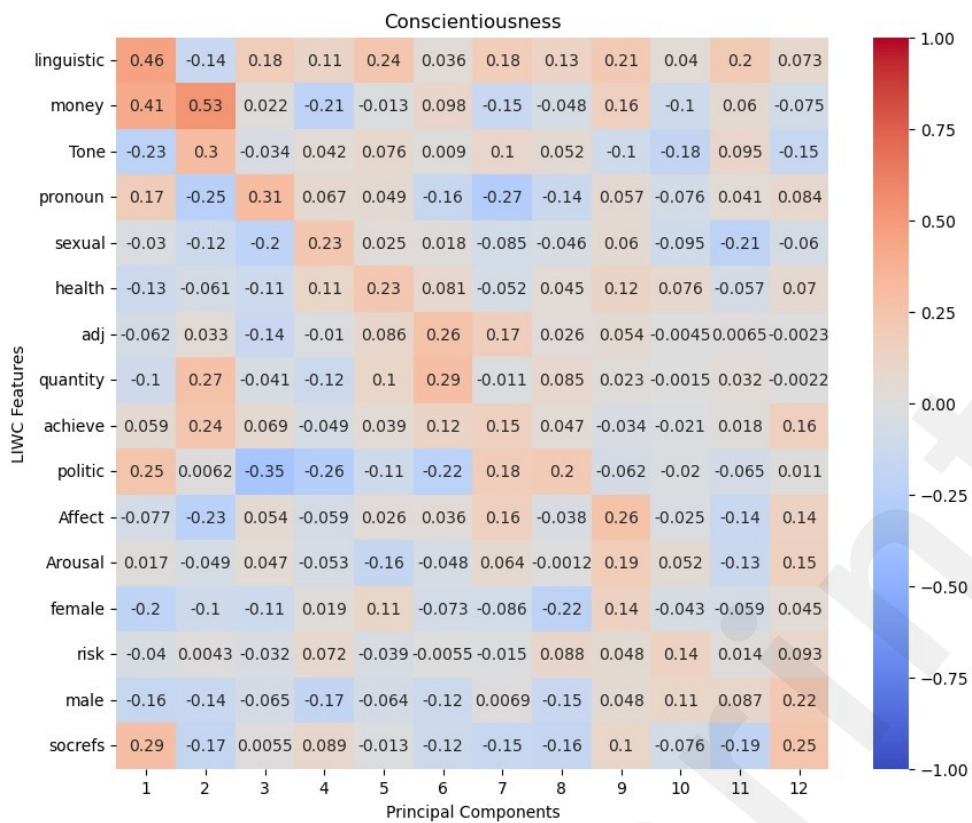
In this section, we explain correlation strategy that we applied on LIWC features and OpenAI embeddings. We begin with generating linguistic features (LIWC) on a test dataset resulted from RQ1. These features from LIWC were highly sparse and redundant. Hence, we run Lasso Regression to retrieve the highly important features for each Personality Traits. Similarly, for the embedding, we performed Principal Component Analysis (PCA) to reduce dimensions. The principal components were iteratively chosen from 5 to 20 to match with the retrieved features from the LIWC.

Figure 3. Correlation score between top 12 LIWC features and OpenAI embedding (reduced PCA components)

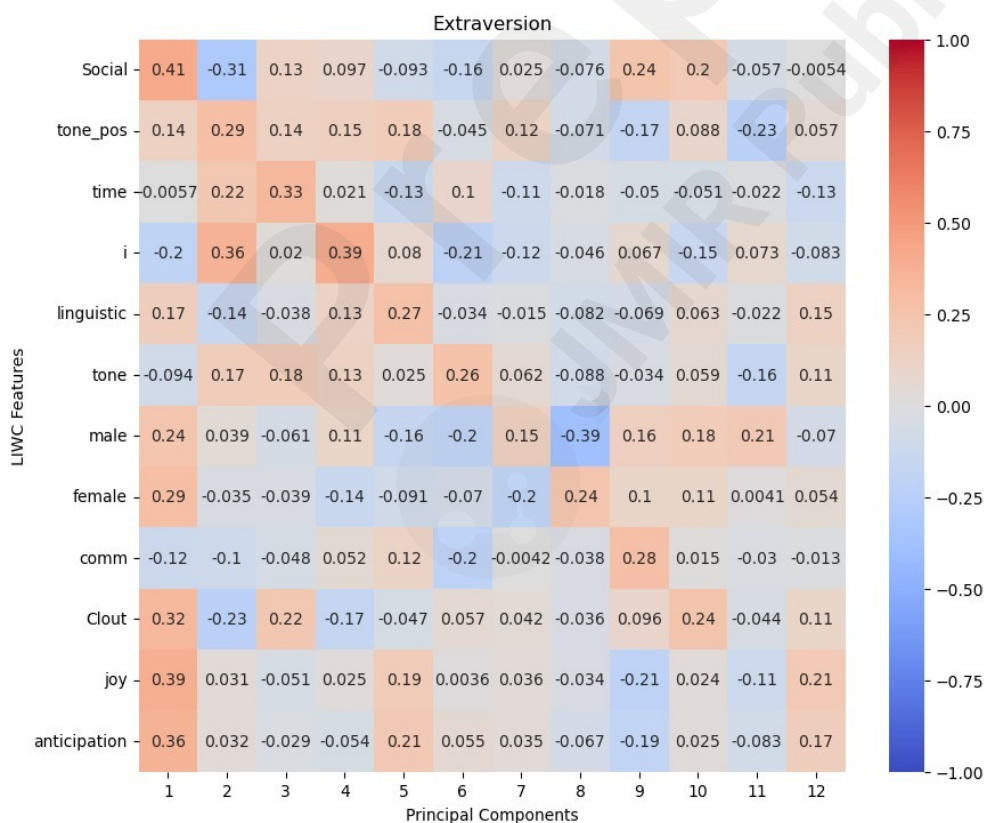
Openness



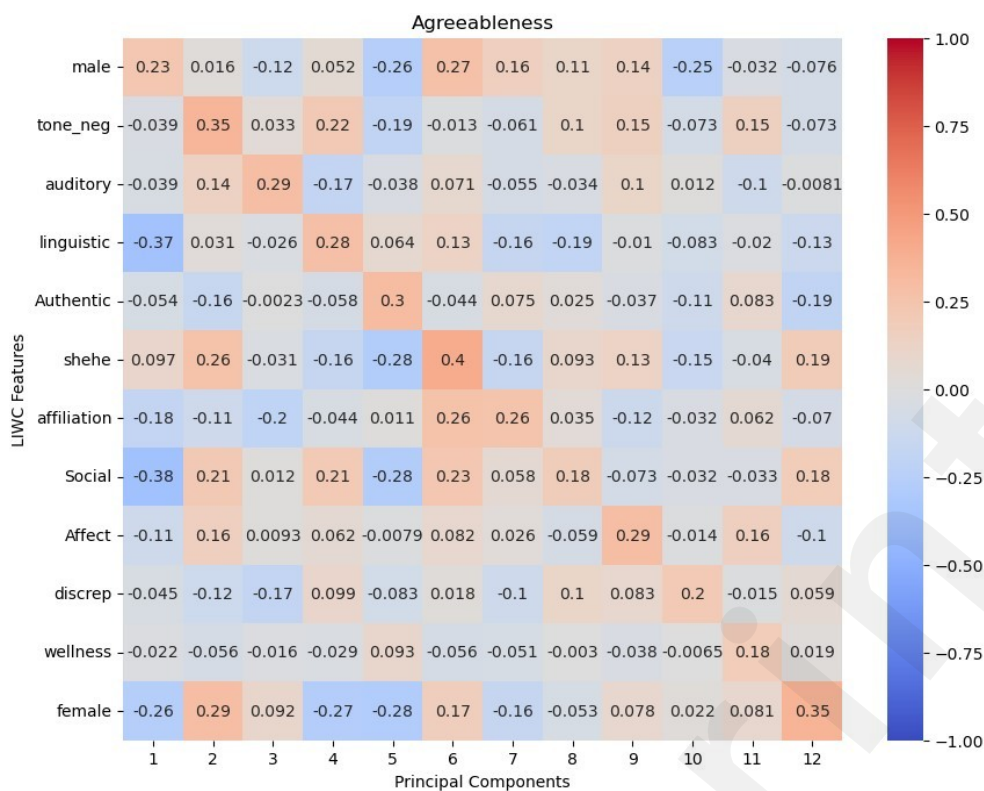
Conscientiousness



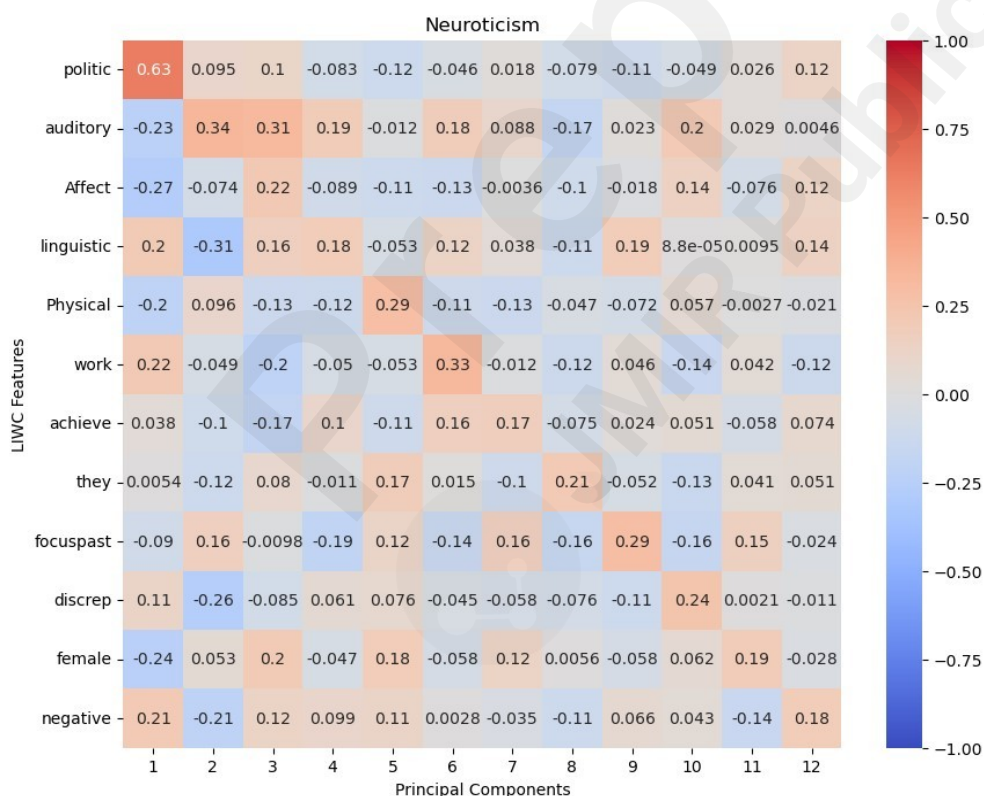
Extraversion



Agreeableness



Neuroticism



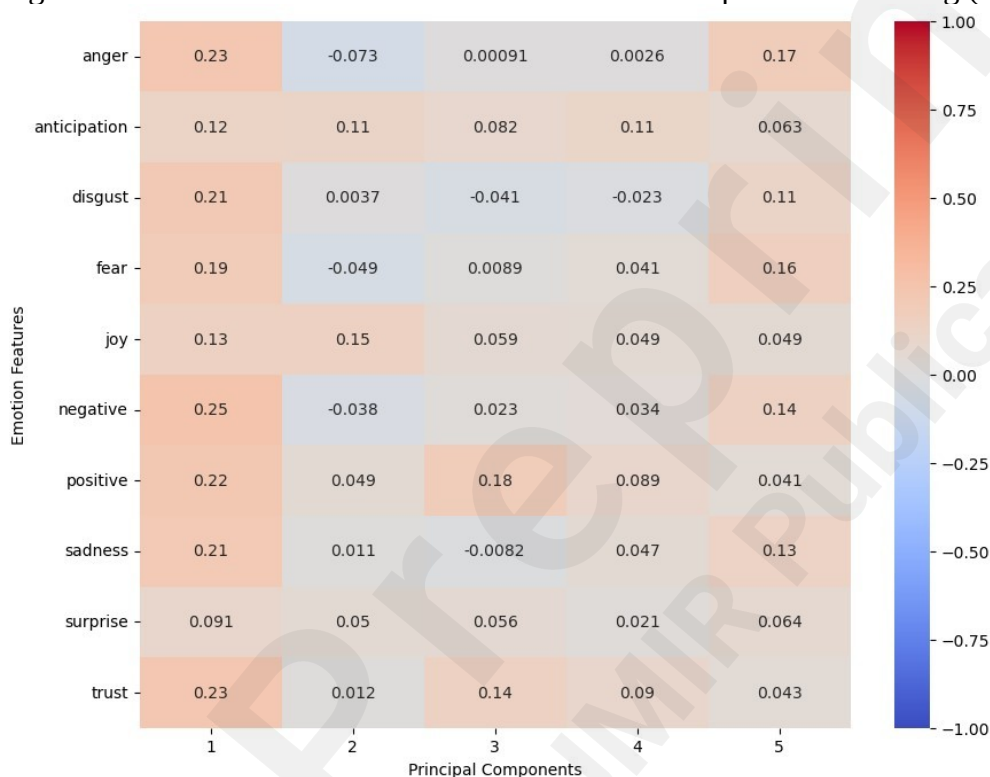
Our analysis from correlation Figure 3 highlights the top 12 LIWC features correlation with the top 12 Principal components across each Personality traits. For the openness and extraversion traits, social feature identified as a key feature with the correlation score of 0.53 and 0.41 respectively. Likewise, linguistic feature has been highlighted as key feature for conscientiousness with a score of 0.46, and politic feature for neuroticism with a score of 0.63. However, for the agreeableness, the

correlation score doesn't account high correlation to any specific feature, although the result shows the high correlation with use of pronouns (she and he, denoted by "shehe") with 0.4 score.

Correlation between Emotion Features and reduced LLM (OpenAI) embedding

In addition to LIWC features, we investigated the emotion features using NRC-Emotion library. For each input sequence in the test dataset, the 10 emotions; anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise and trust were extracted. Similar as with LIWC correlation, we iteratively performed PCA from to reduce embedding dimension from 3 to 10 and evaluated correlation. In the following correlation result, five principal dimensions were observed being correlated with the emotion features.

Figure 4. Correlation between Emotion features and OpenAI embedding (reduced PCA components)



The result from Figure 5 shows that emotion features are captured in only single dimension (first principal axis) rather than multiple dimensions. However, the third dimension highlights the semantic relation between positive emotions, where trust, positive and anticipation were observed with a moderate correlation. Likewise, the negative emotions were observed in the fifth dimension with a moderate correlation among anger, disgust, fear, negative and sadness.

Overall, the correlation with LIWC and emotions features demonstrated that the embeddings essentially capture the key linguistic properties required for personality assessment and provide evidence on the reliability of LLM embeddings for personality prediction.

RQ3. Do other psycholinguistic features, like LIWC, along with feature

engineering help with personality prediction?

In this section, we analyze the impact of psycholinguistic features along with LLMs embedding on the Personality prediction task under four different settings: only linguistic feature, only contextual embedding, both linguistic features and contextual embeddings, and optimal linguistic features (obtained from feature engineering) and contextual embeddings. The performance for this research question is carried out as binary classification task.

Psycholinguistic Features

Along with LIWC features (119 features), we further extracted other features like emotions, sentiments from the tools like NRC Emotion (10 features), NRC VAD (3 features) and VADER Sentiment (1 feature). The total of 133 features contributed as statistical linguistic properties for the experiment.

Feature Engineering

Feature Scaling/Normalization and Imputation

Feature Scaling is the process of transforming the features to a common scale. Scaling features often help the optimization algorithms to converge fast, as the gradient updates will be dominated with larger magnitudes otherwise. Further, scaling also ensures that all the features contribute equally to the model's learning process, preventing bias toward features with larger values. In our work, we used standard scaling with a mean (μ) of 0 and standard deviation (σ) of 1, as $X_{scaled} = \frac{X - \mu}{\sigma}$.

Imputation is the process of handling missing values in the dataset, which leads to biasness or incorrect model training otherwise. We have replaced the missing values with a mean value of the feature column.

Feature Selection

Feature selection is the process of identifying and selecting the most relevant features for building a model such as to speed up training and reducing overfitting. In our work, we experiment with a combination of irrelevant filter, redundant filter, and a standard filter selection method, Mutual Information.

Our combination technique began with an irrelevant filter method to remove filters that have little or no predictive power with respect to the target. For this, we used a Variance threshold method to filter features with very low variance. Then, a redundant filter method was used to remove highly correlated features by using correlation matrix method with different set of thresholds; 0.25, 0.5, 0.6. Finally, we applied the Mutual Information method to measure the information between features and target variables, and select the features based on the ranking on the Mutual Information scores. The top k features selection was selected from the Mutual Information.

Binary Classification and its Evaluation Metric

Dataset

Our comparative analysis under different settings is based on binary classification. Hence, we performed the transformation of the continuous value to discrete target classes (1 and 0). First, we scaled the scores in each of the 3 scoring categories using MinMaxScaler. Then, we compared the scaled score to its quantile value (Median) to divide the scores into 1 and 0. This binarization was carried out in all 3 scoring criteria independently. The distribution of binary classes after transformation is presented in Table 6.

Table 6. Binary Class distribution

Class	Agreeableness	Openness	Conscientiousness	Extraversion	Neuroticism
0	448,041	445,127	424,105	407,798	456,137
1	487,061	489,975	510,997	527,304	478,965

Model

We used the same model architecture as in Question 1 with additional Sigmoid function at the final layer. The sigmoid activation module on the final layer ensured the score between 0 and 1. Simultaneously, we utilized Binary Cross Entropy as a loss function to train the model. For the true output y_i and total input samples N , the loss function is calculated as:

$$BCE(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

Evaluation Metric: AUROC

The Area Under the Receiver Operating Characteristic Curve (AUROC) measures the performance of binary classification models. It quantifies the model's ability to discriminate between positive and negative classes across all possible classification thresholds. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at varying thresholds, and the AUROC represents the area under this curve.

TPR: It measures the proportion of actual positive data points that a model correctly identifies as positive. It is calculated as:

$$TPR = \frac{TP}{TP + TN}$$

FPR: It measures the proportion of negative data points that a model incorrectly identified as positive. It is calculated as:

$$FPR = \frac{FP}{TN + FP}$$

Here, TP and TN refer to the count of data points correctly identified as positive and negative respectively, while FP refers to the count of data points incorrectly identified as positive.

With above TPR and FPR values, AUROC is computed as:

$$AUROC = \int_0^1 TPR(FPR) d(FPR)$$

Generally, AUROC value equal to 1 is considered as a perfect classifier, and value above 0.75 is considered as a good classifier.

Result

Using all above preliminary techniques, we now present the classification performance under four settings as shown in the Table. For all settings, a BiLSTM with attention Model (Model structure same as in RQ1) has been used to train.

Table 7. Binary Classification Performance (AUROC) under 4 different settings

	Only Linguistic Features	Only Contextual (OpenAI) Embedding	All Linguistic Features with Contextual Embedding	Optimal Linguistic Features with Contextual (OpenAI) Embedding
Openness	0.65	0.82	0.64	0.82
Conscientiousness	0.63	0.82	0.62	0.81
Extraversion	0.65	0.83	0.65	0.80
Agreeableness	0.66	0.83	0.66	0.83
Neuroticism	0.65	0.82	0.64	0.81

The result shows the best performance under two settings: Only Contextual Embedding and Optimal Linguistic Features with Contextual Embedding. While feature engineering showed the nearly equal performance as without linguistic features, using all linguistic features, in fact, had a negative impact on the model performance. This highlights the use of contextual embeddings has a greater impact among all.

RQ4. How different LLM embedding works, which one is better?

In the final section of this study, we experiment the performance of embedding from different LLMs; RoBERTa, BERT, OpenAI. These models solely differed from the model size where RoBERTa had 355M, and 175B. Table [8] presents AUROC for the binary classification performance across the LLMs.

Table 8. Binary Classification Performance (AUROC) in different LLMs' embeddings

	RoBERTa	BERT	OpenAI
Openness	0.80	0.74	0.82
Conscientiousness	0.80	0.73	0.82
Extraversion	0.81	0.75	0.83
Agreeableness	0.82	0.73	0.83
Neuroticism	0.80	0.74	0.82

Discussion

In this study, we explored the potential of Large Language Models' (LLMs) embeddings in assessing personality traits using the well-labeled PANDORA Reddit dataset. Our findings demonstrate that LLMs' embeddings effectively capture nuanced psycholinguistic properties essential for personality trait identification, offering a robust alternative to traditional methods. From RQ1, we observed that LLMs' embeddings, when paired with a simple deep learning model (BiLSTM with Attention), significantly outperformed zero-shot inference results. Specifically, conversational AI embeddings, OpenAI, showed superior performance compared to frozen embeddings like BERT and RoBERTa, though the margin of improvement was not substantial. This suggests that while LLMs' embeddings are highly effective, the choice of model architecture plays a nuanced role in performance.

For RQ2, our psychometric validity analysis revealed an average Cronbach's alpha of 0.63, indicating moderate reliability across personality traits. Additionally, the correlation between embeddings and the original dataset's personality traits highlighted the ability of LLMs' embeddings to preserve the underlying structure of the data. Further, dimensionality reduction via PCA and correlation analysis with LIWC and emotion features demonstrated that embeddings effectively encapsulate relevant psycholinguistic information, by showcasing key features like social for extraversion and openness, politic for neuroticism, linguistic for conscientiousness, and pronouns for agreeableness. Also, the polarity of emotion in certain dimension clearly evidenced that embeddings pose the claim.

Regarding RQ3, our experiments with feature engineering revealed that LLMs' embeddings consistently outperformed linguistic features alone. Notably, combining linguistic features with embeddings did not yield significant improvements, suggesting that LLMs' embeddings inherently capture the necessary linguistic and contextual information for personality prediction. This finding highlights the potential of LLMs to reduce reliance on domain specific feature engineering, making them a powerful tool for psychological research.

Finally, in RQ4, we compared the performance of different LLMs' embeddings (RoBERTa, BERT, and OpenAI) and found that OpenAI and RoBERTa embeddings achieved similar scores, with an average AUROC of 0.81 across all personality traits. While OpenAI embeddings slightly outperformed RoBERTa in some cases, the difference was not marginal. Given the computational and financial costs associated with generating OpenAI embeddings, our results suggest that RoBERTa, as an open-source model, provides a cost-effective and equivalent alternative for personality prediction tasks. This finding highlights the practical utility of open-source models like RoBERTa, especially in resource-constrained settings, without compromising on performance.

Overall, our results underscore the efficacy of LLMs' embeddings in personality prediction and their ability to generalize across diverse psycholinguistic contexts. The consistent performance of embeddings highlights their potential as a leading tool for psychological and personality-related research, while also emphasizing the importance of model selection in achieving optimal results.

Limitation and Future Work

Despite the promising results, our study has several limitations that warrant consideration. First, the moderate validity observed in RQ2, with an average Cronbach's alpha of 0.63, restricts our ability to make definitive validation inferences. While the embeddings demonstrated reliability, the moderate correlation suggests room for improvement. Future work could explore alternative validation methods, such as confirmatory factor analysis (CFA) or item response theory (IRT), to further assess the psychometric properties of LLMs' embeddings. Second, while OpenAI embeddings showed

superior performance in RQ4, the study did not include other advanced conversational AI agents, such as DeepseekAI, as competitors. Including these models could provide a more comprehensive comparison and strengthen the generalizability of our findings. Additionally, the computational and financial costs associated with OpenAI embeddings raise practical concerns, making open-source models like RoBERTa a more viable option for many applications. However, further exploration of cost-performance trade-offs is needed to guide model selection in real-world scenarios. Third, our experiments were conducted on a single dataset (PANDORA), which may limit the generalizability of our findings. Applying the learned models to other datasets, such as MyPersonality or Twitter-based datasets, could provide additional insights into the robustness and transferability of LLMs' embeddings across different domains and contexts. Addressing these limitations in future research could significantly improve the confidence and applicability of LLMs' embeddings in psychological and personality-related studies, paving the way for more reliable and scalable solutions in this domain.

Conclusion

This study highlighted the effectiveness of LLMs' embeddings in predicting personality traits, with OpenAI and RoBERTa embeddings demonstrating superior performance over zero-shot inference and traditional linguistic feature engineering methods. While OpenAI embeddings slightly outperformed RoBERTa, the cost trade off makes RoBERTa viable alternative. The moderate reliability of embeddings and their strong correlation with linguistic features underscore their potential for psychological research. However, limitations such as dataset specific generalizability and the exclusion of other advanced conversational AI agents indicate areas for future exploration. Overall, LLMs' embeddings present a scalable and efficient approach to personality prediction, with significant implications for advancing psychological assessments.

Abbreviation

AI Artificial Intelligence

LLM Large Language Model

LIWC Linguistic Inquiry and Word Count

BERT Bidirectional Encoder Representations from Transformers

RoBERTa Robustly Optimized BERT Approach

BiLSTM Bidirectional Long Short-Term Memory

PCA Principal Component Analysis

References

1. Murray, H. A. (1936). Basic concepts for a psychology of personality. *The Journal of general psychology*, 15(2), 241-268.
2. Goodstein, L. D., & Lanyon, R. I. (1999). Applications of personality assessment to the workplace: A review. *Journal of business and psychology*, 13, 291-322.
3. Costa Jr, P. T., & McCrae, R. R. (2010). The five-factor model, five-factor theory, and interpersonal psychology. *Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions*, 91-104.
4. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018.
5. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*

arXiv:2501.12948.

6. Gjurković, M., Karan, M., Vukojević, I., Bošnjak, M., & Šnajder, J. (2020). PANDORA talks: Personality and demographics on Reddit. *arXiv preprint arXiv:2004.04460*.
7. Lin, H. (2024). DLP-personality detection: a text-based personality detection framework with psycholinguistic features and pre-trained features. *Multimedia Tools and Applications*, 83(13), 37275-37294.
8. Killian Jr, J., & Sun, R. (2024, July). Detecting Big-5 Personality Dimensions from Text Based on Large Language Models. In *International Conference on Deep Learning Theory and Applications* (pp. 264-278). Cham: Springer Nature Switzerland.
9. Sirasapalli, J. J., & Malla, R. M. (2023). A deep learning approach to text-based personality prediction using multiple data sources mapping. *Neural Computing and Applications*, 35(28), 20619-20630.
10. Christian, H., Suhartono, D., Chowanda, A., & Zamli, K. Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8(1), 68.
11. Yang, K., Lau, R. Y., & Abbasi, A. (2023). Getting personal: A deep learning artifact for text-based measurement of personality. *Information Systems Research*, 34(1), 194-222.
12. Chen, Y. (2023). *Exploring the Intercorrelations of Big Five Personality Traits: Comparing Questionnaire-Based Methods and Automated Personality Assessment using BERT and RNN Models* (Master's thesis).
13. El-Demerdash, K., El-Khoribi, R. A., Shoman, M. A. I., & Abdou, S. (2022). Deep learning based fusion strategies for personality prediction. *Egyptian Informatics Journal*, 23(1), 47-53.
14. Radisavljević D, Rzepka R, Araki K. Personality Types and Traits—Examining and Leveraging the Relationship between Different Personality Models for Mutual Prediction. *Applied Sciences*. 2023; 13(7):4506. <https://doi.org/10.3390/app13074506>
15. Peters H, Matz SC. Large language models can infer psychological dispositions of social media users. *PNAS Nexus*. 2024;3(6):pgae231. Published 2024 Jun 13. doi:10.1093/pnasnexus/pgae231].
16. Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 466, 10622–10643.
17. Wang, Y., Zhao, J., Ones, D. S., He, L., & Xu, X. (2025). Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1), 519. <https://doi.org/10.1038/s41598-024-84109-5>
18. Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2023). PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. *NAACL-HLT*.
19. Safdari, M., Serapio-Garc'ia, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., & Matari'c, M. (2023). Personality Traits in Large Language Models. *ArXiv*, abs/2307.00184.
20. Huang, M., Zhang, X., Soto, C., & Evans, J. (2024). Designing LLM-Agents with Personalities: A Psychometric Approach. *ArXiv*, abs/2410.19238.].
21. Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1), 117
22. Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001), 2001.
23. Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30, 457-500.
24. Mohammad, S. M., & Turney, P. D. (2013). Nrc emotion lexicon. *National Research Council, Canada*, 2, 234.

25. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. **J Lang Soc Psychol.** 2010;29(1):24-54.
26. Boyd RL, Pennebaker JW. Language-based personality: A new approach to personality in a digital world. **Curr Dir Psychol Sci.** 2017;26(4):373-380.
27. Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9), e73791.
28. Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6), 934.
29. Salsabila, G.D. and Setiawan, E.B. 2021. Semantic Approach for Big Five Personality Prediction on Twitter. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*. 5, 4 (Aug. 2021), 680 - 687. DOI:<https://doi.org/10.29207/resti.v5i4.3197>.
30. Plank, B., & Hovy, D. (2015, September). Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 92-98).
31. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
32. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
33. Oz, H. (2015). Personality traits and ideal L2 self as predictors of academic achievement among prospective English teachers. In *ICERI2015 Proceedings* (pp. 5833-5841). IATED.
34. GitHub repository: PersonaClassifier. <https://github.com/JulinaM/PersonaClassifier>

Supplementary Files