A Systematic Evaluation of Transformer-LM Representations for Capturing Author States and Traits

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used in human-centered applications, yet their ability to model diverse psychological constructs is not well understood. In this study, we systematically evaluate a range of Transformer-based LLMs to predict psychological variables across five major dimensions: affect, substance use, mental health, sociodemographics, and personality. Analyses span three temporal levels—short daily text responses, two-week, and user-level text collected over two years—allowing us to examine how each model's strengths align with the underlying stability of different constructs. The findings show that mental health signals emerge as the most reliably captured dimension, possibly because people often use detailed, specific language when describing their emotional experiences, which makes these cues easier for models to detect. At the daily scale, context-rich embeddings of DeBERTa and HaRT excel at capturing short-term emotional fluctuations, whereas few-shot Llama3-8B proves particularly adept at modeling nuanced substance use behaviors at the two-week interval. Aggregating text over the entire study period yields stronger correlations for socio-demographic factors e.g., age, income. These results suggest actionable insights into the design of LLM-based approaches for psychological assessments, emphasizing the importance of selecting appropriate model architectures and temporal aggregation techniques to the stability and nature of the target construct.

1 Introduction

Recently, LM representations (that is, embeddings) have shown strong promise in improving psychological assessments of mental health and well-being now approaching the theoretical upper limit in accuracy for some outcomes (Kjell et al., 2022). However, their utility in different constructs is inconsistent. A systematic evaluation is yet to be performed to determine what types of psychological attributes can best be captured in language (Boyd and Markowitz, 2024), and by which LM. Psychological variables differ by many factors, fundamentally including (a) their *stability* – from being more *state*-like (i.e. changing frequently) to more *trait*-like (i.e. changing slowly) as well as (b) their *construct domains* – encompassing areas such as emotional states, personality traits, cognitive functions, and behavioral tendencies.

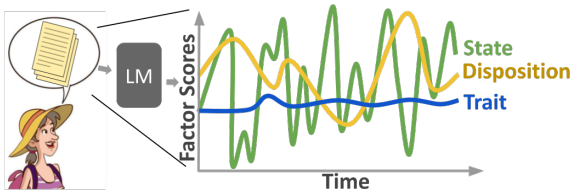

Figure 1: Conceptual framework illustrating how Language Models (LMs) capture temporal dynamics of psychological constructs across varying levels of stability. Constructs are categorized into states (highly variable, e.g., mood), dispositions (moderately stable, e.g., stress), and traits (highly stable, e.g., personality). This figure underscores the study's focus on aligning LM architectures with psychological stability to enhance predictive performance across temporal granularities.

In this study, we systematically evaluate the effectiveness of Transformer-LM representations for capturing human psychological states and traits from textual data. Unlike prior work that has predominantly focused on inferring the mental states of conversational participants (Jara-Ettinger and Rubio-Fernandez, 2021), our approach centers on self-expressed narratives—treating participants as the authors of their own psychological profiles. By leveraging a unique dataset consisting of language collected in bursts over daily intervals – Ecological Momentary Assessments (EMA) – and aggregated over longer periods (wave-level) across a two-year span, we directly compare the capability of various LLM-based embeddings to predict

standardized psychological scores. These scores span multiple domains, including affect/emotion, personality, mental health, sociodemographics, and health behaviors, thereby offering a comprehensive assessment of language models in psychological evaluation. (Nilsson et al., 2024)

Our key contributions are: (1) We provide a systematic comparison of different LLM representations' ability to capture 43 psychological variables; (2) We characterize LLMs by the domains they capture and their ability to capture more stable to less stable attributes; (3) We introduce a method based on measurement theory to determine the stability of psychological constructs, characterizing the temporal granularity at which it is best measured for downstream analysis; (4) We evaluate different outcome aggregation methods to assess their impact on capturing psychological constructs accurately; and finally, (5) We offer best practices for leveraging LLMs in psychology-related tasks, recommending which models and aggregation strategies are best suited for different types of variables.

2 Related Work

Understanding psychological states and traits through language has been a longstanding focus of both psychology and computational linguistics. The evolution of analyzing psychological states through language has progressed from simple lexicon-based tools such as LIWC (Boyd et al., 2022) to more sophisticated approaches using distributed word embeddings (Carducci et al., 2018) and ultimately to advanced contextualized models for capturing personality (Safdari et al., 2023), emotions (Al-Twairesh, 2021), and substance use (Shah-Mohammadi and Finkelstein, 2024). Concurrently, researchers have examined stable individual differences by initially employing bag-of-words and topic modeling methods to predict personality traits, gender, and socioeconomic status (Schwartz et al., 2013).

Modern transformer-based architectures have enhanced capturing the nuanced relationships between words and their broader context, enabling more accurate psychological modeling through language analysis (Kjell et al., 2022). Encoder-based models have emerged as a popular choice for modeling psychological constructs within natural language processing (NLP) since their bidirectional nature helps capture richer semantics (Qorib et al., 2024; Saattrup Nielsen et al., 2024), with significantly fewer parameters than decoder-only counterparts. They have been applied to identify various psychological factors ranging from mental health (Greco et al., 2023; Bauer et al., 2024), cognitive styles (Zong et al., 2020; Varadarajan et al., 2024b), affective factors (Salmerón-Ríos et al., 2024; Li et al., 2024; Hasan et al., 2024; Fabunmi et al., 2025; Xu and Jiang, 2024), substance behaviors (Mahbub et al., 2025) to stable factors like personality (Lynn et al., 2020; Santos and Paraboni, 2022) and demographic attributes (Benton et al., 2016). However, previous studies have also demonstrated the effectiveness of other types of model architectures - human-level LMs like HaRT (Soni et al., 2022) in capturing subtle psychological changes in text (Ganesan et al., 2022); auto-regressive LMs like GPT-2 for assessing stable traits like personality (Soni et al., 2024); prompting-based methods have also been successfully applied to identify language patterns indicative of mental health, personality factors (V Ganesan et al., 2023; Varadarajan et al., 2024a). In this work we focus on the modern LLM architectures given the high degree of variability of architectures in state-of-the-art mental health and psychology-related tasks. Unlike prior studies, we assess LLMs performance across multiple temporal granularities (daily messages, biweekly waves, and aggregated user histories) using a longitudinal dataset.

3 Dataset

The dataset for this study was collected by Nilsson et al. (2024), by recruiting service industry workers in the United States, who are often underrepresented in NLP research. The participant population is socioeconomically diverse and predominantly comprises middle-aged individuals who identify as female, representing approximately 75% of the participants. Their ages range from 18 to 68 years, with a mean age of 35 years (standard deviation: 9.47). These individuals were recruited through professional and online groups within the United States and are known to be at high risk for excessive alcohol use (Jose et al., 2022). This demographic choice is particularly relevant for studies focusing on occupational stressors prevalent within the service industry.

Over a two-year period, data were collected in six 14-day "waves" via a custom smartphone application. The participants were prompted three times a day for Ecological Momentary Assess-

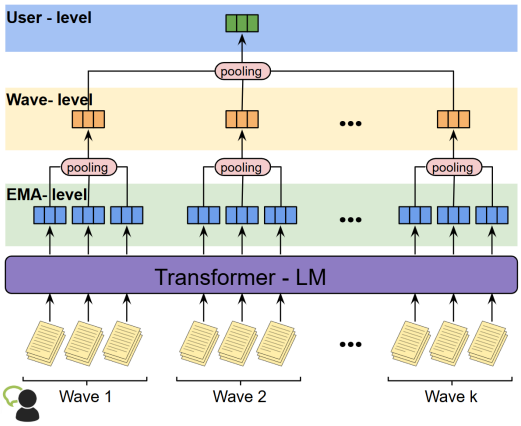

Figure 2: Hierarchical Embedding Generation Model for Psychological State and Trait Analysis. This diagram depicts the process of generating embeddings at different levels: EMA-level(message-level), wave-level, and user-level, using a large language model (LLM).

Variables	EMA (~daily) ICC	EMA (~daily) retest	Wave (~quarterly) ICC	Wave (~quarterly) retest
Arousal (ARO)	.108	.173	.443	.490
Valence (VAL)	.295	.313	.712	.724
No of Drinks (DRI)	.391	.423	.776	.738
Craving (CRA)	.412	.481	.607	.726
Daily Stress(PSS1)	.547	.622	.776	.738
Stress(PSS)	-	-	.580	.586
AUDIT C (AUC)	-	-	.710	.698
Positive Affect (PAF)	-	-	.668	.710
Negative Affect (NAF)	-	-	.602	.642
Openness (OPE)	-	-	.680	.660
Conscientiousness (CON)	-	-	.693	.682
Extraversion (EXT)	-	-	.778	.810
Agreeableness (AGR)	-	-	.638	.617
Neuroticism (NEU)	-	-	.747	.767
GAD7 (GAD)	-	-	.720	.736
PHQ9 (PHQ)	-	-	.753	.739
MACE (MAC)	-	-	.775	.775
Individual Income (INC)	-	-	.768	.793
Age (AGE)	-	-	.995	.997

Table 1: Stability Metrics of Psychological Variables at EMA and Wave Levels. This table reports two key stability measures for each variable: the intra-class correlation coefficient (ICC) and the average test-retest Pearson correlation (retest). Higher values indicate greater stability over time (i.e., less fluctuation between assessments). Cells are color-coded to denote standard stability categories: > 0.7: High (blue); 0.5 to 0.7: Medium (yellow); < 0.5: Low (green) (Koo and Li, 2016). Variables with dashes were not assessed at the EMA level.

ments (EMAs), where they provided a textual response ($>$ 200 words) in English to the prompt: *"Please describe in 2 to 3 sentences how you are currently feeling."* Each wave thus yielded up to 42 responses per participant, totaling a potential 42 x 6 = 252 responses across all six waves. To ensure data quality, only individuals who completed at least two waves with a minimum of two responses per wave were included, resulting in 10,108 EMAs from 120 participants across 406 total user waves. Alongside textual responses, an EMA also consisted of ratings for affect, stress level, number of drinks and craving for alcohol. Additionally,at the start of each wave, participants completed comprehensive questionnaires evaluating personality traits, mental health status, affective states, stress, and alcohol use. This design integrates dynamic, day-to-day EMA data with baseline wave-level assessments, offering a holistic longitudinal perspective on participants' static characteristics and evolving emotional and behavioral patterns.

4 Methodology

This research explores three hierarchical levels of data analysis: the **message-level**, which focuses on individual EMA responses; the **wave-level**, which involves analysis of aggregated responses over 14-day periods; and the **user-level**, which aggregates data across all waves to examine long-term psychological trends.

Models. We systematically evaluate a broad set of Transformer-based architectures autoencoder models—such as BERT (Devlin, 2018), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020); encoder-decoder architectures like T5 (Raffel et al., 2020) and FLAN-T5 (Chung et al., 2022); as well as autoregressive models including GPT-2 (Radford et al., 2019), HaRT (Soni et al., 2022) - specifically designed for human language modeling, XLNet (Yang et al., 2019), Llama2 (Touvron et al., 2023b), and Llama3 (Touvron et al., 2023a). We further examined the influence of architectural scaling by incorporating both base and large variants of these models. We also benchmark Llama3-8B-Instruct via zero-shot and few-shots prompting, comparing its performance against embedding-based pipelines to capture psychological constructs.

Outcomes. In this research, we assessed a broad set of psychological variables spanning five dimensions: **Affect** such as *Valence* and *Arousal* (Remington et al., 2000), and *Positive and Negative Affect (PANAS)* (Thompson, 2007); **Substance behavior** like *Number of Drinks*, *AUDIT-C* (Bradley et al., 2007), *Craving*, and *MACE* (Lange et al., 2017); **Mental health** indicators including *PHQ9* for depression (Kroenke et al., 2003), *GAD7* for

Variable Group: Model \ (stability)	States (low)		Dispositions (mid)		Traits (high)		User Level	
	r	MSE	r	MSE	r	MSE	r	MSE
Auto Encoder								
RoBERTa-base	0.36**	3.10	0.39	7.98	0.40	30.49	0.39	19.93
RoBERTa-large	0.38**	3.04	0.37	8.07	0.39	30.85	0.39	19.37
BERT-base	0.35**	3.11	0.36	8.54	0.38	32.57	0.39	20.19
DeBERTa-base	0.37**	3.06	0.35	8.12	0.42	29.90	0.41	19.01
DeBERTa-large	**0.39****	3.00	0.39	8.27	0.41	30.70	0.43	19.44
Encoder-Decoder								
T5-large	0.37**	3.03	0.38	8.05	0.39	29.97	0.44	18.79
FLAN T5-large	0.38**	2.96	0.37	8.20	0.38	30.88	0.40	19.29
AutoRegressive								
GPT2-medium	0.36**	3.04	0.36	8.15	0.40	30.07	0.41	18.98
GPT2 HLC	0.36**	3.04	0.36	8.09	0.41	30.61	0.41	20.19
Xlnet-large	**0.39****	2.90	0.38	7.87	0.42	29.20	0.45	18.07
Llama2-7B	0.35*	3.04	0.36	8.52	0.37	31.97	0.37	20.66
Llama3-8B	0.34*	3.03	0.33	8.68	0.36	31.86	0.34	20.29
Human LM								
HaRT	0.34*	2.92	0.39**	7.65	0.44	28.13	0.46	17.91
Zero Shot								
Llama3-8B	0.28	4.48	0.35	28.14	0.44	42.12	0.43	32.29
Few Shot								
Llama3-8B	0.35	4.08	**0.40**	23.55	**0.48**	37.79	**0.48**	29.05

Table 2: Accuracy (as average Pearson r) and Mean Squared Error (MSE) of model embeddings for capturing states, dispositions, and traits. *States* are variables with low stability (high variability) across time, while *dispositions* have moderate stability and *traits* have high stability. Statistically significant differences from zero-shot Llama3-8B baseline: * ($p < .05$) and ** ($p < .001$), computed using boot-strapped resampling across individuals over 1000 trials. Models tend to perform progressively better in predicting more stable attributes.

anxiety (Plummer et al., 2016), *general stress (PSS)* (Cohen et al., 1983), and *daily stress (PSS Nervous Stress Agreement)*; **Socio-demographics** such as *Income* and *Age*; and **Personality** traits like *Openness*, *Neuroticism*, *Conscientiousness*, *Extraversion*, and *Agreeableness* from the BIG-5 model (Soto and Jackson, 2013). These variables were selected to cover a broad range of psychological states and traits, allowing for a precise evaluation of LLMs' ability to model these constructs effectively. All outcomes were self-reported. *Valence*, *Arousal*, *Number of Drinks*, *Craving*, and *Daily Stress* were collected at the EMA level, while the remaining variables were obtained once per wave.

Stability. To assess the stability of psychological measures across different temporal dimensions, this study utilized two statistical metrics:

1. Intra-class Correlation Coefficient (ICC) (Liljequist et al., 2019) is calculated to assess the reliability of measurements by quantifying the proportion of total variance attributable to differences between individuals. The total variance of a variable is decomposed into two components: the between-individual variance ($\sigma^2_{\text{between}}$), representing variability in measurements across different individuals, and the within-individual variance (σ^2_{within}), capturing variations in repeated measurements for the same individual over time. The ICC is computed as:

$$\text{ICC} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}$$

Higher ICC values indicate greater consistency and reliability of measurements over time.

2. Test-Retest Reliability Each variable in the dataset was analyzed to assess the consistency of measurements across multiple time points. A Pearson correlation matrix was constructed to evaluate pairwise correlations between all waves for each variable. To isolate inter-wave correlations, the lower triangular portion of the matrix (excluding the diagonal) was extracted. The mean of these inter-wave correlations was then calculated, providing a single summary metric for each variable to quantify its *test-retest reliability* (Weir, 2005).

The test-retest reliability metric is computed as:

$$\text{retest} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r_{ij}$$

where n: the number of temporal points; and r_{ij}: the Pearson correlation coefficient between temporal points i and j.

Although ICC scores provide insight into inter-individual stability, our stability classification focused on test-retest scores, as they more directly reflect temporal consistency and are better suited for categorizing constructs along the state-disposition-trait continuum.

Table 1 presents the stability metrics assessed at both EMA-level and Wave-level. Attributes such as *Age* and *Personality Traits* (e.g., Extraversion, Conscientiousness) demonstrate high stability, with ICCs and test-retest correlation coefficients (r) exceeding 0.7, indicating their enduring nature. Conversely, moderately stable measures like *Stress* (PSS) and *Positive Affect* display stability indices between 0.5 and 0.7. More dynamic metrics, including *Valence*, *Arousal*, and *Craving*, show lower stability (< 0.5), reflecting their sensitivity to immediate environmental and situational changes. These results illustrate the need to customize modeling strategies according to the temporal stability of each metric. High-stability metrics benefit from

Dimensions	Aff	Sub	Mnt	SDe	Per
Auto Encoder					
RoBERTa-base	.487**	.220	.600	.346	.302
RoBERTa-large	.489**	.222	.581	.349	.296
BERT-base	.499**	.172	.585	.333	.296
DeBERTa-base	.467**	.233	.603	.406	.300
Deberta-large	.509**	.261	.590	.425	.317
Encoder-Decoder					
T5-large	.502**	.229	.595	.421	.309
FLAN T5-large	.495**	.195	.604	.355	.297
Auto Regressive					
GPT2-medium	.490**	.248	.595	.339	.289
GPT-2$_{HLC}$	.501**	.246	.601	.323	.278
Xlnet-large	.503**	.285	.599	**.439**	.315
Llama2-7B	.459*	.168	.586	.308	.299
Llama3-8B	.456*	.132	.567	.261	.268
Human LM					
HaRT	**.512***	.321	**.615****	.395	**.342**
Zero Shot Prompting					
Llama3-8B	.456	.397	.535	.315	.285
Few Shot Prompting					
Llama3-8B	.492	**.446**	.597	.382	.324

Table 3: Performance Evaluation of LMs across different dimensions of psychology. Statistically significant differences from zero-shot Llama3-8B baseline: * (p < .05) and ** (p < .001), computed using boot-strapped resampling across individuals over 1000 trials. (**Aff** stands for affective variables; **Sub** for substance behavior variables; **Mnt** stands for mental health variables; **Sde** stands for socio-demographics and **Per** stands for personality). Similar trends are observed across models, with mental health dimensions being captured more effectively than all other dimensions.

Parameter	VAL	ARO	PSS1	DRI	CRA
N	**10,108**	**10,108**	**4,638**	**8,185**	**4,909**
AutoEncoder					
RoBERTa-base	.624*	.373**	.562*	.211	.246
RoBERTa-large	.635*	.389**	.530	.260	.242
BERT-base	.602	.354**	.545	.178	.249
DeBERTa-base	.624*	.375**	.554	.222	.277
DeBERTa-large	**.648****	**.404****	.562	.285	.218
Encoder-Decoder					
T5-large	.633**	.390**	.556	.244	.239
FLAN T5-large	.642**	.392**	.573*	.259	.281
AutoRegressive					
GPT2-medium	.602	.355**	.524	.226	.250
GPT-2$_{HLC}$	.615	.358**	.542	.222	.239
XLNet-large	.627*	.392**	.545	.271	.281
LLama2-7B	.596	.365*	.529	.228	.210
LLama3-8B	.588	.361*	.490	.225	.214
Human LM					
HaRT	.632*	.331*	**.583****	**.296**	**.317**
Zero Shot Prompting					
Llama3-8B	.470	.161	.377	.225	.245
Few Shot Prompting					
Llama3-8B	.606	.315	.531	.273	.261

Table 4: Pearson Correlation Coefficients for EMA Level Analysis. Highlighted cells indicate which model excels for the corresponding outcome. Statistically significant differences from zero-shot Llama3-8B baseline: * ($p < .05$) and ** ($p < .001$), computed using boot-strapped resampling across individuals over 1000 trials. Overall, EMA-level analysis highlights how certain architectures excel at capturing fleeting emotional cues, while others better align with daily behavioral variables. Although prompting can enhance performance to some extent, many embedding-based models still provide a stronger and more consistent signal.

aggregation methods across waves or individuals, enhancing model accuracy. Conversely, metrics with low stability require approaches that can capture transient, moment-specific dynamics.

Experimental Design. We generated average token embeddings from each LLM to predict psychological outcomes, employing a 10-fold cross-validation scheme with ridge regression—selected for its ability to address multicollinearity and mitigate overfitting. Data stratification was applied to ensure balanced outcome distributions across folds. Model performance was then assessed using Pearson correlation coefficient (r) and Mean Square Error (MSE). Figure 2 illustrates our multi-level embedding aggregation strategy. We first generated token embeddings for each EMA message, capturing immediate linguistic and emotional cues. Next, we aggregated these embeddings at the wave level (i.e., over each 14-day period) to capture mid-range patterns within users. Finally, all wave-level embeddings were averaged at the user level, creating a holistic profile of each participant over the full study duration. This progressive approach enables a nuanced examination of daily, mid-range, and long-term psychological dynamics through the corresponding text signals.

Computational Framework. The computational framework leveraged PyTorch, HuggingFace, and the Differential Language Analysis Toolkit (DLATK) (Schwartz et al., 2017) for feature extraction and model evaluation. To support these operations, 2 NVIDIA RTX A6000 GPUs with 48GB of VRAM, were used for generating embeddings and executing prompting tasks.

5 Results

Capturing States, Dispositions, and Traits. Table 2 presents the average Pearson correlation coefficients (r) that measure the accuracy of vari-

Dimension	Affective				Substance Behavior				Mental Health				Personality					SocioDemog.	
Variable	**VAL**	**ARO**	**PAF**	**NAF**	**DRI**	**CRA**	**AUC**	**MAC**	**GAD**	**PHQ**	**PSS**	**PSS1**	**OPE**	**CON**	**EXT**	**AGR**	**NEU**	**INC**	**AGE**
N	**406**	**406**	**133**	**133**	**406**	**179**	**406**	**126**	**406**	**179**	**406**	**406**	**345**	**345**	**345**	**345**	**345**	**406**	**132**
Auto Encoder																			
RoBERTa-base	.769	.360	.364	.551	.247	.398	.316	-.017	**.550**	.731	.498	.550	.130	.279	.187	.322	.464	.252	.371
RoBERTa-large	**.797**	.351	.353	.577	.261	.259	.381	-.038	.538	.677	.495	.477	.098	.254	.236	.238	**.538**	.240	.355
BERT-base	.746	.361	.347	.556	.201	.305	.290	-.081	.511	.732	.480	.488	.081	.268	.267	.251	.487	.268	.334
DeBERTa-base	.780	.337	.348	.348	.251	.459	.293	-.019	.545	**.740**	.490	.516	.151	.259	.252	.273	.480	.323	.382
DeBERTa-large	.777	**.398**	.326	.555	.309	.288	.346	.037	.526	.692	.493	.506	.181	.268	.234	.302	.520*	.352	.403
Encoder-Decoder																			
T5-large	.773	.343	.292	.550	.243	.292	.302	-.027	.542	.714	.462	.508	.057	.314	.217	.314	.446	.324	.430
FLAN T5-large	.764	.344	.313	.568	.228	.336	.338	-.124	**.550**	.705	.488*	.535	.066	.273	.279	.251	.411	.244	.352
Auto Regressive																			
GPT2-medium	.770	.366	.369	.579	.244	.418	.283	.025	.525	.709	.503	.518	.018	.300	.231	.276	.452	.226	.372
GPT2 HLC	.772	.352	**.433**	.576	.269	.479	.320	-.039	.538	.722	.487	.514	-.029	.268	.256	.271	.466	.287	.290
Xlnet-large	.788	.386	.308	.566	.322	.355	.344	.110	.547	.664	.503	.522	.127	.294	.273	.270	.474	.302	.427
Llama2-7B	.780	.345	.272	.502	.223	.319	.250	.050	.539	.712	.440	.505	.122	.247	.117	.315	.507	.261	.271
Llama3-8B	.785	.291	.319	.491	.190	.338	.248	-.048	.543	.697	.436	.466	.066	.180	.110	**.323**	.526	.253	.229
Human LM																			
HaRT	.771	.243	.410	**.646**	.339	.346	.387	.228	.536	.732	.495	.563	**.188**	.277	.252	.304	.495	**.366***	.317
Zero Shot																			
Llama3-8B	.753	.292	.414	.505	.351	.470	.471	.336	.495	.527	.442	.660	.050	.291	.201	.296	.428	.190	.426
Few Shot																			
Llama3-8B	.768	.294	.426	.510	**.400**	**.498**	**.473**	**.477**	.493	.552	**.533**	**.696**	.055	**.400***	**.322**	.308	.443	.203	**.531**

Table 5: Pearson Correlation Coefficients for Wave-Level Analysis. Highlighted cells indicate which model excels for the corresponding outcome. Higher values indicate stronger predictive ability. Statistically significant differences from zero-shot Llama3-8B baseline: * (p < .05) and ** (p < .001), computed using boot-strapped resampling across individuals over 1000 trials. Different model families appear stronger for capturing mental health signals through embeddings, while prompting approaches are more effective in modeling substance behaviors. In contrast, personality traits remain less directly inferred from two-week text bursts, yielding only moderate correlations overall.

ous LLMs in capturing three distinct categories of psychological constructs: states, dispositions, and traits. *States* exhibit low stability with high variability, reflecting transient conditions like momentary emotions or acute stress. *Dispositions* show moderate stability, capturing semi-consistent attributes such as enduring moods or habitual behaviors. *Traits*, on the other hand, are highly stable and represent consistent characteristics, such as core personality traits. In the autoencoder category, DeBERTa-large embeddings achieved the highest correlations across all stability levels— r=0.39 for states, r=0.39 for dispositions, and r=0.41 for traits—demonstrating its effectiveness in capturing both fleeting emotional cues and more stable psychological characteristics. In the encoder-decoder family, T5-large and flanT5- large embeddings show competitive performance, indicating that these models are balanced in capturing both short-term and enduring psychological signals. For the autoregressive models, XLNet-large embeddings stands out, achieving an r=0.39 for states, r=0.38 for dispositions, and r=0.42 for traits, with the highest user-level correlation r=0.45. The human-aware model HaRT also showed strong performance for stable constructs, highlighting the benefits of its hierarchical attention mechanism. The zero-shot configuration of Llama3-8B demonstrates relatively low performance for states r=0.28 but improves for traits r=0.44 and user-level measures r=0.43. However, with few-shot prompting using between 2 to 7 examples per outcome Llama3-8B shows marked improvement across the board, with correlations rising to r=0.35 for states, r=0.40 for dispositions, and notably r=0.48 for traits and user-level outcomes. This improvement suggests that minimal task-specific prompts can significantly enhance the model's ability to capture subtle psychological signals, particularly for more stable constructs. Across all models, there is a clear trend: constructs with greater stability—such as traits and user-level aggregates—are predicted with higher accuracy than dynamic, state-like variables. These results imply that language signals associated with enduring characteristics are more consistent and easier to model, whereas the fleeting nature of transient states poses greater challenges.

Performance Evaluation of LMs across various psychological dimensions. Table 3 evaluates the effectiveness of LLMs in capturing five psychological dimensions—Affective, Substance Behavior, Mental Health, Socio-Demographics, and Personality—using average Pearson correlation coefficients. The results highlight distinct strengths among architectures. However, a common pattern emerges across nearly all models, with **mental**

Dimension	Affective				Substance Behaviour				Mental Health				Personality					SocioDemog.	
Variable	**VAL**	**ARO**	**PAF**	**NAF**	**DRI**	**CRA**	**AUC**	**MAC**	**GAD**	**PHQ**	**PSS**	**PSS1**	**OPE**	**CON**	**EXT**	**AGR**	**NEU**	**INC**	**AGE**
N	**120**	**120**	**103**	**103**	**120**	**94**	**120**	**99**	**120**	**120**	**120**	**92**	**120**	**120**	**120**	**120**	**120**	**120**	**103**
Auto Encoder																			
RoBERTa-base	.769	.148	.371	.543	.218	.383	.123	.073	.571	.587	.648	.706	.031	.369	.257	.409	.569	.398	.364
RoBERTa-large	.778	.142	.332	.538	.260	.390	.173	.033	.587	.578	.659	.691	.055	.343	.258	.357	.585	.393	.406
BERT-base	.758	.344	.388	.534	.179	.359	.034	.007	.579	.598	.634	.702	.046	.360	.226	.367	.611	.384	.346
DeBERTa-base	.768	.202	.340	.550	.207	.406	.180	.054	.590	.591	.662	.739	.045	.390	.278	.372	.502	.454*	.467
DeBERTa-large	.766	.309	.369	.539	.304	.433	.301	.085	.573	.592	.657	.709	.068	.370	.254	.383	.589	**.470***	.477
Encoder-Decoder																			
T5-large	.773	.346	.361	**.559**	.215	.449	.220	.112	.565	.601	.658	**.749**	.172	.374	.271	.402	.525	.421	.508
FLAN T5-large	.771	.277	.339	.539	.172	.432	.141	-.114	.592	.605	.648	.736	.166	.325	.284	.395	.523	.448*	.377
Auto Regressive																			
GPT2-medium	.767	.199	.355	.533	.243	.431	.224	.139	.601	.614	.660	.703	.002	.388	.247	.400	.580	.396	.361
GPT2 HLC	**.784**	.218	.390	.514	.280	.478	.234	-.020	.599	.604	.670	.733	-.019	.360	.242	.380	.589	.440*	.274
Xlnet-large	.766	.328	.333	.536	.331	.398	.309	.128	.598	.613	**.688**	.707	.084	.367	.317	.399	.548	.440	**.586**
Llama2-7B	.764	.184	.290	.494	.053	.335	-.007	.024	.580	.611	.644	.714	.076	.368	.237	.417	.579	.378	.323
Llama3-8B	.771	.242	.269	.441	.040	.341	-.053	-.174	.565	.579	.637	.694	-.061	.353	.163	.375	**.650**	.343	.219
Human LM																			
HaRT	.774	**.377**	.410	.531	.313	.483	.282	.218	.604	**.630**	.680	.708	**.174**	.393	**.320**	.424	.592	.447*	.450
Zero Shot																			
Llama3-8B	.734	.174	.512	.550	.427	.473	.568*	.408	.562	.564	.570	.620	.065	.377	.234	.419	.486	.135	.507
Few Shot																			
Llama3-8B	.741	.184	**.528**	.546	**.453**	**.521**	**.611**	**.492**	**.739**	.609	.583	.640	.093	**.424**	.251	**.440**	.501	.228	.565

Table 6: Pearson Correlation Coefficients for User-Level Analysis. Highlighted cells indicate which model excels for the corresponding outcome. Higher values indicate stronger predictive ability. Statistically significant differences from zero-shot Llama3-8B baseline: * ($p < .05$) and ** ($p < .001$), computed using boot-strapped resampling across individuals over 1000 trials. When aggregating all participant text at the user level, personality traits see modest gains but remain moderate overall, reflecting that everyday writing rarely includes explicit statements about one's personality. Socio-demographic outcomes (age, income) show improved correlations when there is enough text to reveal contextual signals.

health variables generally yielding stronger correlations than the other four dimensions. This suggests that language-based signals for mental health may be more readily detected or more consistently expressed in the text, enabling LLMs to track these constructs more effectively.

Predictive Performance Across Temporal Granularities. Figure 3 illustrates trends in the predictive performance of models for *Valence*, *Arousal*, and *Stress* across different temporal granularities. For *Valence*, performance is consistently low across all models at the EMA level, reflecting the challenge of capturing this construct in momentary assessments. However, predictive accuracy improves markedly at the wave and user levels, indicating that temporal aggregation enhances stability and predictive reliability for this construct. In contrast, *Arousal* demonstrates an inverse trend, with higher predictive performance at the EMA level due to its immediate and dynamic nature. As temporal aggregation progresses to wave and user levels, performance diminishes, highlighting the difficulty of capturing this transient construct in aggregated representations. For *Stress*, the predictive performance shows a balanced progression across all levels, with models performing moderately at the EMA level and improving at the wave and user levels. This indicates that Stress encompasses both dynamic and stable components, benefiting from temporal aggregation to capture broader patterns while retaining its sensitivity to momentary fluctuations.

Longitudinal Analysis of Model Performance. Across the three temporal granularity levels—EMA, wave, and user—our findings reveal notable shifts in which models excel and how reliably they can capture each psychological construct (Tables 4, 5, and 6). At the **EMA level**, where text mirrors daily, state-like fluctuations (e.g., valence, arousal, or sudden stress spikes), context-rich embeddings such as those from DeBERTa-large and HaRT are especially effective (Table 4). These architectures appear to capture subtle linguistic indicators tied to momentary emotional swings and daily behaviors, reflected in their high Pearson correlations. By contrast, when aggregating data over two-week intervals (**wave level**), we model more disposition-like constructs (e.g., short-term stress, negative affect) that are neither purely momentary nor completely stable. Here, most embedding-based approaches continue to perform well (RoBERTa, DeBERTa), whereas Llama3-8B few-shot prompting often proves advantageous, particularly for capturing **substance use behaviors** like alcohol intake and craving (Table 5). This suggests that targeted prompting can highlight context-specific cues about consumption patterns

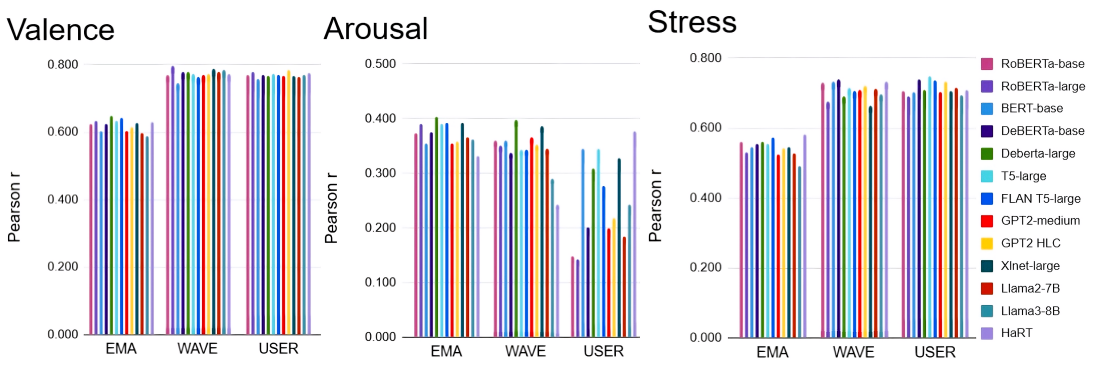

Figure 3: The predictive performance of different transformer-LM models across varying temporal granularities—EMA, Wave, and User. Valence and stress are more accurately predicted at the wave or user level while arousal has greater accuracy at the EMA (i.e. document) level.

or cravings, thereby boosting performance beyond what embedding-only pipelines achieve. Nonetheless, personality-related signals remain harder to extract from these mid-length narratives, producing only moderate correlations overall—likely because users' short-term text lacks the introspective detail needed to reveal deeper trait-like characteristics. Finally, at the **user level** (Table 6), which spans up to two years of participants' collected text, the gains in predicting personality remain modest, reflecting how everyday narratives not fully include explicit trait-related content. Meanwhile, **sociodemographic factors** such as age and income exhibit clearer signals when enough text is aggregated to reveal contextual mentions—references to life events, employment histories, or financial concerns. Across Tables 5 and 6, it is also evident that outcomes with sparse data or low prevalence sometimes register near-zero or slightly negative correlations (refer to Appendix figure A1 for an illustration of the effect of sample size).

Overall, these results emphasize the importance of aligning model architecture and aggregation strategies with the stability and nature of psychological dimensions across temporal levels. Additionally, while encoder-based models are well accepted to be the best representation, our results indicate that this is not always the case—other architectures and prompting methods can outperform them, depending on the psychological construct and its temporal stability.

6 Conclusion

This study systematically evaluated the capabilities of many Transformer-based LLMs for capturing human factors across different levels of temporal stability—low, medium, and high as well as different psychological dimensions. The findings reveal that the model performance is highly influenced by the temporal granularity of the data, the stability of the outcomes, and the constructs being modeled. While aggregation strategies proved instrumental for enhancing predictive reliability for stable constructs (Traits), low-stability constructs that undergo a lot of dynamic fluctuations on the daily might not be best represented through averages over time, an effect we specifically observe for some *States*.

Additionally, we introduce a framework that determines the preferred temporal granularities at which these constructs should be analyzed. This framework not only improves the modeling of psychological constructs, but also has practical implications for data collection design in future studies. By identifying the optimal data collection frequency for such experiments, it offers the potential to eliminate the need for costly daily surveys when evaluating LLMs' capabilities in psychological assessment. Together, these insights emphasize the importance of tailoring model selection, data aggregation strategies, and experimental design to align with the unique temporal characteristics and stability of psychological constructs, paving the way for more efficient and reliable LLM-based approaches to psychological evaluation.

7 Limitations

This study has few limitations that stem from the ecological nature of its design. The participant population, although socioeconomically diverse, was restricted to service industry workers who spoke English as the primary language recruited from pro-

fessional organizations and online groups within the United States. The resulting population represents the service industry, consisting of 75% females. The mean age of the participants is 35 years, with a standard deviation of 9.47 years. The ages range from a minimum of 18 to a maximum of 68 years. The population is known to be at high risk for excessive alcohol use (Jose et al., 2022). Thus, the results found here should not be assumed to generalize to other populations.

Furthermore, limitations exist related to the scope of LLMs and the outcomes assessed. Relative to the largest LLMs, the models used here were smaller ($\sim$ 8 billion parameters), so findings may not generalize to larger models. We suggest that further work should test this for those with the fortune of such computational resources. Our study specifically aimed to evaluate a diverse set of model architectures and types (autoencoders, encoder-decoders, autoregressive) at a tractable scale to provide insights into their comparative performance across temporal granularities and psychological constructs. Although this study uniquely assessed embeddings across a wide range of specific and general psychological outcomes, it certainly does not represent the whole range of possible psychological factors. Future research should expand these findings to additional psychological variables such as in the cognitive domain to enhance their applicability.

8 Ethical Considerations

This study adhered to rigorous ethical guidelines to ensure the responsible application of artificial intelligence in mental health research with procedures approved by an independent academic Institutional Review Board (IRB). All participants provided informed consent for the use of their data in this study, with no agreement to share their non-anonymized individual data beyond the scope of this research. This work is aimed at advancing interdisciplinary NLP-psychology research to better understand human behaviors as reflected in language. Importantly, the models and methods developed in this study are not intended or validated for deployment in clinical settings or for other commercial applications, such as targeted marketing. Instead, the focus is on contributing to the development of more accurate and ethically sound techniques that benefit society and promote human health.

References

Nora Al-Twairesh. 2021. The evolution of language models applied to emotion analysis of arabic tweets. *Information*, 12(2):84.

Brian Bauer, Raquel Norel, Alex Leow, Zad Abi Rached, Bo Wen, and Guillermo Cecchi. 2024. Using large language models to understand suicidality in a social media–based taxonomy of mental health disorders: Linguistic analysis of reddit posts. *JMIR mental health*, 11:e57234.

Adrian Benton, Raman Arora, and Mark Dredze. 2016. Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19.

Ryan L. Boyd and David M. Markowitz. 2024. Verbal behavior and the future of social science. *American Psychologist*, pages 1–23. Place: US Publisher: American Psychological Association.

Ryan L Boyd et al. 2022. The development and psychometric properties of liwc-22. *Behavior Research Methods*, 54:2502–2520.

Katharine A Bradley, Anna F DeBenedetti, Robert J Volk, Emily C Williams, Danielle Frank, and Daniel R Kivlahan. 2007. Audit-c as a brief screen for alcohol misuse in primary care. *Alcoholism: Clinical and Experimental Research*, 31(7):1208–1217.

Giulio Carducci, Giuseppe Rizzo, Diego Monti, Enrico Palumbo, and Maurizio Morisio. 2018. Twitpersonality: Computing personality traits from tweets using word embeddings and supervised learning. *Information*, 9(5):127.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4):385–396.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Oluwatoba Fabunmi, Saman Halgamuge, Daniel Beck, and Katja Holtta-Otto. 2025. Large language models for predicting empathic accuracy between a designer and a user. *Journal of Mechanical Design*, 147(4).

Adithya V Ganesan, Vasudha Varadarajan, Juhi Mittal, Shashanka Subrahmanya, Matthew Matero, Nikita Soni, Sharath Chandra Guntuku, Johannes Eichstaedt, and H Andrew Schwartz. 2022. Wwbp-sqt-lite: Multi-level models and difference embeddings for moments of change identification in mental health forums. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 251–258.

Candida M Greco, Andrea Simeri, Andrea Tagarelli, and Ester Zumpano. 2023. Transformer-based language models for mental health issues: a survey. *Pattern Recognition Letters*, 167:204–211.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024. Llm-gem: Large language model-guided prediction of people's empathy levels towards newspaper article. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2215–2231.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Julian Jara-Ettinger and Paula Rubio-Fernandez. 2021. Quantitative mental state attributions in language understanding. *Science advances*, 7(47):eabj0970.

Rupa Jose, Matthew Matero, Garrick Sherman, Brenda Curtis, Salvatore Giorgi, Hansen Andrew Schwartz, and Lyle H. Ungar. 2022. Using facebook language to predict and describe excessive alcohol use. *Alcoholism: Clinical and Experimental Research*, 46(5):836–847.

Oscar NE Kjell, Sverker Sikström, Katarina Kjell, and H Andrew Schwartz. 2022. Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports*, 12(1):3918.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2003. The patient health questionnaire-2: Validity of a two-item depression screener. *Medical Care*, 41(11):1284–1292.

Shannon Lange, Charlotte Probst, Kevin D. Gmel, Jürgen Rehm, and Svetlana Popova. 2017. Worldwide prevalence of fetal alcohol spectrum disorders: A systematic literature review including meta-analysis. *Addiction*, 112(9):1520–1532.

Ming Li, Yusheng Su, Hsiu-Yuan Huang, Jiali Cheng, Xin Hu, Xinmiao Zhang, Huadong Wang, Yujia Qin, Xiaozhi Wang, Kristen A Lindquist, et al. 2024. Language-specific representation of emotion-concept knowledge causally supports emotion inference. *iScience*, 27(12).

David Liljequist, Björn Elfving, and Kirsten Skavberg Roaldsen. 2019. Intraclass correlation–a discussion and demonstration of basic features. *PloS one*, 14(7):e0219854.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Veronica Lynn, Niranjan Balasubramanian, and H Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5306–5316.

Maria Mahbub, Gregory M Dams, Sudarshan Srinivasan, Caitlin Rizy, Ioana Danciu, Jodie Trafton, and Kathryn Knight. 2025. Decoding substance use disorder severity from clinical notes using a large language model. *npj Mental Health Research*, 4(1):5.

August Håkan Nilsson, Hansen Andrew Schwartz, Richard N Rosenthal, James R McKay, Huy Vu, Young-Min Cho, Syeda Mahwish, Adithya V Ganesan, and Lyle Ungar. 2024. Language-based ema assessments help understand problematic alcohol consumption. *Plos one*, 19(3):e0298300.

Faye Plummer, Laura Manea, Dominic Trepel, and Dean McMillan. 2016. Screening for anxiety disorders with the gad-7 and gad-2: a systematic review and diagnostic metaanalysis. *General Hospital Psychiatry*, 39:24–31.

Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are decoder-only language models better than encoder-only language models in understanding word meaning? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16339–16347.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Technical Report.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nicholas A. Remmington, Leandre R. Fabrigar, and Penny S. Visser. 2000. Reexamining the circumplex model of affect. *Journal of Personality and Social Psychology*, 79(2):286–300.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2024. Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual nlu tasks. *arXiv e-prints*, pages arXiv–2406.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Alejandro Salmerón-Ríos, José Antonio García-Díaz, Ronghao Pan, and Rafael Valencia-García. 2024. Fine grain emotion analysis in spanish using linguistic features and transformers. *PeerJ Computer Science*, 10:e1992.

Vitor Garcia dos Santos and Ivandré Paraboni. 2022. Myers-briggs personality classification from social media text using pre-trained language models. *arXiv preprint arXiv:2207.04476*.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, S Ramones, Megha Agrawal, Achal Shah, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. In *PloS one*, volume 8.

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Johannes Eichstaedt, and Lyle Ungar. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60. Association for Computational Linguistics.

Fatemeh Shah-Mohammadi and Joseph Finkelstein. 2024. Extraction of substance use information from clinical notes: Generative pretrained transformer–based investigation. *JMIR Medical Informatics*, 12:e56243.

Nikita Soni, Niranjan Balasubramanian, H Schwartz, and Dirk Hovy. 2024. Comparing pre-trained human language models: Is it better with human context as groups, individual traits, or both? In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 316–328.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Christopher J Soto and Joshua J Jackson. 2013. Five-factor model of personality. *Journal of Research in Personality*, 42:1285–1302.

Edmund R. Thompson. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of Cross-Cultural Psychology*, 38(2):227–242.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023a. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Pierre Albert, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H. Andrew Schwartz. 2023. 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*, pages 390–400, Toronto, Canada. Association for Computational Linguistics.

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, et al. 2024a. Archetypes and entropy: Theory-driven extraction of evidence for suicide risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 278–291.

Vasudha Varadarajan, Syeda Mahwish, Xiaoran Liu, Julia Buffolino, Christian Luhmann, Ryan L Boyd, and H. Andrew Schwartz. 2024b. Capturing Human Cognitive Styles with Language: Towards an Experimental Evaluation Paradigm. In *To Appear in North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL 2025*.

Joseph P Weir. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *Journal of Strength and Conditioning Research*, 19(1):231–240.

Zhichao Xu and Jiepu Jiang. 2024. Multi-dimensional evaluation of empathetic dialog responses. *arXiv preprint arXiv:2402.11409*.

Zhilin Yang, Zihang Dai, Yiming Yang, et al. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, pages 5753–5763.

Shi Zong, Alan Ritter, and Eduard Hovy. 2020. Measuring forecasting skill from text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5317–5331, Online. Association for Computational Linguistics.

A Appendix

A.1 Zero - shot Prompting

This prompt template was provided to the model without any additional in-context examples, allowing it to generate responses solely based on the given instructions and the input text.

```
<|begin_of_text|>
<|start_header_id|>
system
<|end_header_id|>
```

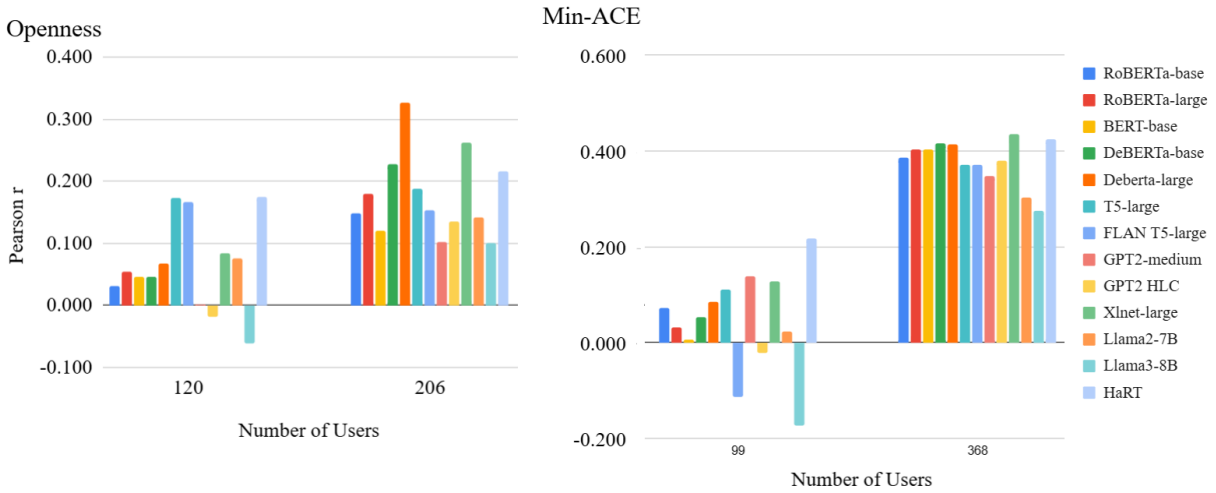

Figure A1: Model Performance vs Sample Size for Selected Outcomes. This figure illustrates how the predictive performance of different LLM-based embeddings (represented by distinct colored bars) varies as the number of users increases. For smaller sample sizes (shown on the left side of each plot), correlations often hover near zero or even become slightly negative, reflecting instability and potential noise due to insufficient data. As the number of users increases, most models' correlations tend to improve, highlighting that more data generally leads to more reliable embeddings and more accurate predictions of psychological constructs.

```
You are helpful assistant.
<|eot_id|>
<|start_header_id|>
user
<|end_header_id|>
{prompt}
Provide your assessment by
responding with "Score: "
followed
by the corresponding number.
<|eot_id|>
<|start_header_id|>
assistant
<|end_header_id|>
```

A.2 Few - shot Prompting

This prompt template was provided to the model along with 2 to 7 in-context examples from the dataset for different outcomes, enabling it to generate responses by leveraging both the given instructions and the provided examples in addition to the input text.

```
<|begin_of_text|>
<|start_header_id|>
system
<|end_header_id|>
You are helpful assistant.
<|eot_id|>
<|start_header_id|>
user
<|end_header_id|>
{prompt}
Provide your assessment by
responding with "Score: "
followed by a single integer
considering the following
examples:
{few_shot_examples}.
<|eot_id|>
<|start_header_id|>
assistant
<|end_header_id|>
```

Tables A1 - A6 contains the prompts designed for each psychological variable analyzed in this study. These prompts were tailored to elicit meaningful responses from the language models (LLMs) by framing the tasks in a clear, context-specific manner. While these prompts provide a solid starting point, they can be further refined to enhance clarity and alignment with the constructs of interest, thereby improving the reliability and generalizability of zero-shot and few shot prompting methodologies.

VARIABLE	PROMPT
Valence	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the emotional valence of the writer based by analyzing their essays. The valence scale measures the degree of pleasantness or unpleasantness, with 0 representing very low levels of pleasantness and 4 representing very high levels of pleasantness.
Arousal	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the arousal of the writer by analyzing their essays. The arousal scale measures the energy of the writer, with 0 representing calm or lethargic energy, and 2 representing active or excited energy.
Positive Affect	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the Positive Affect score of the writer by analyzing their message. Positive Affect (from PANAS) refers to the extent to which an individual experiences positive emotional states such as being interested, excited, enthusiastic, proud, or inspired. Look for explicit mentions of these emotions, descriptions of situations that evoke positive feelings, and implicit cues in the tone, choice of words, or overall mood conveyed in the text. Estimate the Positive Affect score experienced by the writer with a number between 5-25, with 5 representing low positive affect, while 25 denotes very high positive affect.
Negative Affect	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the Negative Affect score of the writer by analyzing their essays. Negative Affect (from PANAS) refers to the extent to which an individual feels negative emotional states such as distress, fear, anger, guilt, or nervousness. Look for explicit mentions of emotions, descriptions of situations that might evoke negative feelings, and implicit cues in the tone, choice of words, or overall mood conveyed in the text. Estimate the Negative Affect score experienced by the writer with a number between 5-25, with 5 representing low negative affect, while 25 denotes very high negative affect.

Table A1: Affective Variables

VARIABLE	PROMPT
Individual Income	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the income of the writer based by analyzing their essays. Use your judgment to evaluate references to occupation, lifestyle, education, and financial indicators mentioned in the text. Chose the most closes income range of this individual from the following categories. (A) <$10,000 (B) $10,000-$20,000 (C) $20,000-$30,000 (D) $30,000-$40,000 (E) $40,000-$50,000 (F) $50,000-$60,000 (G) $60,000-$70,000 (H) $70,000-$80,000 (I) $90,000-$100,000 (J) >$100,000. Choose only one option from the above, and respond with "Income Category: ", followed by the alphabet indicative of the corresponding income range.
Age	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the age of the writer based by analyzing their essays, which can range from 18 to 65. Based on linguistic patterns, cultural references, mentions of life events, maturity of the writing and the overall tone, estimate the age of the writer.

Table A2: Socio demographics Variables

VARIABLE	PROMPT
PSS1	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the nervousness/stress levels of the writer by analyzing their essays. The level of stress ranges from 1 to 5 where 1 means very low or no stress/nervousness, and 5 means extremely high stress/nervousness.
PSS	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the severity of stress experienced by the writer based off their essays. Note that stressed individuals are overwhelmed by difficulties in their lives, while individuals who are not stressed are confident in solving their personal problems. Estimate the stress level of the writer based on the Percieved Stress Scale with a number between 0-16, with 0 representing no stress and 16 representing very high levels of stress.

Table A3: Stress Variables

VARIABLE	PROMPT
GAD7	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the anxiety levels experienced by the writer based off their essays. Note that anxious individuals feel nervous, worry too much about different things, have trouble relaxing, or can be easuuly annoyed. Estimate the anxiety level of the writer based on the Generalized Anxiety Disorder scale with a number between 0-21, with 0 representing no anxiety and 21 representing high levels of anxiety.
PHQ9	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the depression levels experienced by the writer based off their essays. Note that depressed individuals feel hopeless, have little interest in doing everyday things, suffer somatic symptoms like abnormal sleep, abnormal appetite, fatigue or psychomotor agitation/retardation. They might also experience trouble with concentrating on things, feelings of worthlessness/guilt or suicidal ideation. Estimate the depression severity of the writer based on the Patient Health Questionnaire scale with a number between 0-27, with 0 representing no anxiety and 27 representing high levels of depression.

Table A4: Mental Health Variables

VARIABLE	PROMPT
No of Drinks	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to analyze the essays and determine the likely number of alcoholic drinks the author of these essays had consumed. Consider any direct mentions of drinking, contextual hints about social settings, behaviors associated with drinking and any indirect references that may imply the consumption of alcohol. Use your expertise to gauge the number of drinks based on the narrative provided.
AUDITC	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to estimate the individual's level of alcohol use based on their essays. Pay close attention to any direct mentions of drinking, contextual clues about social settings, behaviors commonly associated with alcohol consumption, and indirect references that suggest the frequency and quantity of drinking. Assign a score between 0 and 12 using the AUDIT-C scale, where 0 indicates no alcohol use or minimal risk, and 12 indicates a high risk for harmful drinking behaviors or potential alcohol use disorder.
Craving	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to analyze the essays and and determine the intensity of the author's craving for alcohol. Use your judgment to analyze descriptions of feelings, situations triggering desire, any direct mentions of wanting to consume alcohol, or behaviors associated with drinking. Based on the essays, determine how strong the craving is on a scale from 0 to 10, where 0 indicates no craving at all and 10 indicates an extremely high craving.
MACE	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to estimate the individual's level of alcohol cravings based on their message. Pay close attention to any strong urges to drink, descriptions of picturing alcohol or drinking, mentions of imagining the taste of alcohol, reflections on how the body might feel after drinking, and intrusive thoughts about alcohol. Assign a score between 0 and 50 using the Mini-ACE scale, where 0 indicates no cravings or minimal risk, and 50 indicates a high level of persistent cravings or potential risk for harmful drinking behaviors.

Table A5: Substance Behavior Variables

VARIABLE	PROMPT
Openness	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the openness score of the writer by analyzing their essays. Note that individuals who are open to experiences tend to be intellectual, imaginative, sensitive and open-minded while individuals that are not open to experiences tend to be down to earth, insensitive and conventional. The openness scale ranges from 0 to 20, where 0 indicates very low levels of openness and 20 indicates extremely high levels of openness .
Conscientiousness	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the conscientiousness score of the writer by analyzing their essays. Note that individuals who are conscientious tend to be careful, thorough, organized and scrupulous while individuals that are not conscientious tend to be irresponsible, disorganized and unscrupulous. The conscientiousness scale ranges from 0 to 20, where 0 indicates very low levels of conscientiousness and 20 indicates extremely high levels of conscientiousness .
Extraversion	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the extraversion score of the writer by analyzing their essays. Note that individuals who are extraverted tend to be sociable, talkative, assertive and active while individuals that are not extraverted tend to be retiring, reserved and cautious. The conscientiousness scale ranges from 0 to 20, where 0 indicates very low levels of extraversion and 20 indicates extremely high levels of extraversion .
Agreeableness	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the agreeableness of the writer based by analyzing their essays. Note that individuals who are agreeable tend to be good-natured, compliant, modest, gentle, and cooperative while individuals that are not agreeable tend to be irritable, ruthless, suspicious and inflexible. The agreeableness scale ranges from 0 to 20, where 0 indicates very low levels of agreeableness and 20 indicates extremely high levels of agreeableness.
Neuroticism	Carefully read the series of essays posted below, written by a person describing how they felt each day. Each day's essay is separated by a new line, with the most recent one in the bottom. Essays: message Your task is to determine the neuroticism score of the writer by analyzing their essays. Note that individuals who are neurotic tend to be anxious, depressed, angry and insecure while individuals that are not neurotic tend to be calm, poised and emotionally stable. The neuroticism scale ranges from 0 to 20, where 0 indicates very low levels of neuroticism and 20 indicates extremely high levels of neuroticism .

Table A6: BIG-5 Personality Traits Variables