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

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used for human-centered tasks, yet their ability to model diverse psychological constructs is not well understood. This study systematically evaluates the capabilities of diverse Transformer-based LLMs in modeling human psychological constructs across varying levels of temporal stability. Using a unique dataset of Ecological Momentary Assessments (EMAs) at varying levels of aggregation from none (EMA-level) to waves (quarterly), and users (averaged over $\sim$ 2 years), we explore how autoencoder, encoder-decoder, and autoregressive models capture traits and states. The findings reveal that the performance of LLMs is influenced by the level of analysis, with models excelling at specific combinations of outcome stability and construct characteristics. Aggregation strategies play a critical role in enhancing the reliability of predictions for rapidly changing *states*, moderately stable *dispositions*, and enduring *traits*. 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.

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., 2023). 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.

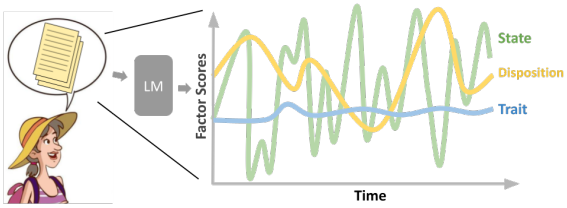

Figure 1: Conceptual framework illustrating how Language Models (LMs) capture temporal dynamics of psychological constructs across varying levels of stability. Constructs are categorized into states (highly variable, e.g., mood), dispositions (moderately stable, e.g., stress), and traits (highly stable, 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 evaluated how well open LM-based representations capture human psychological states and traits from textual data. Using a unique dataset of language captured in bursts over days ("EMA-level") at multiple times (wave level) over the course of 2 years, we compare the ability of LLM-based representations to predict psychological scores from standard questionnaires covering the domains of affect/emotion, personality, mental health, sociodemographics, and health behaviors (Nilsson et al., 2024). We empirically compare three categories of LLMs: autoencoders, encoder-decoder, and autoregressive models, as well as the impact of data aggregation methods across temporal resolutions (e.g., EMAs within days, waves over months, and users overall).

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) evaluation of outcomes aggregation methods to capture psychological constructs; and finally, (5) we list best practices suggested by the results for one to effectively leverage LLMs in psychology-related tasks, including which models are best suited for which types of variables and aggregation strategies.

2 Related Work

Understanding psychological states and traits through language has been a longstanding focus of both psychology and computational linguistics. Traditional approaches primarily relied on lexicon-based tools, such as the Linguistic Inquiry and Word Count (LIWC) tool (Boyd et al., 2022), which maps word usage to psychological categories. LIWC has been widely adopted for psychological assessments (Tausczik and Pennebaker, 2010), providing insights into personality, emotional states, and social behaviors.

Historically, studies on language-psychosocial connections employed traditional methods like bag-of-words and lexicon-based approaches to predict psychological variables, but these lacked the capacity to capture the nuanced and contextual nature of language necessary for modeling complex psychological states. The advent of Large Language Models (LLMs), with their ability to generate contextual embeddings, has addressed these limitations by encoding deeper semantic and syntactic information (Liu et al., 2019; Yang et al., 2019).

Machine learning techniques introduced statistical models capable of identifying linguistic patterns linked to psychological constructs. Early works employed topic modeling (Resnik et al., 2015) and n-gram features (Schwartz et al., 2013a,b) to predict mental health conditions and personality traits. These methods, while valuable, often lacked the depth needed to capture intricate psychological signals embedded in language.

The advent of word embeddings, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), represented a significant advancement by providing dense vector representations of words, improving the modeling of semantic relationships. Despite their success in psychological modeling tasks (Preoţiuc-Pietro et al., 2015), these embeddings lacked contextual awareness, treating each word independently of its surrounding text.

Contextualized word embeddings, introduced through models like ELMo (Peters et al., 2018) and transformer-based architectures such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-2 (Radford et al., 2019), marked a transformative leap. These models consider the context in which words appear, enabling a nuanced understanding of psychological language. Recent studies have leveraged such models for mental health assessment (Coppersmith et al., 2018), personality prediction (Mehta et al., 2020), and detecting psychological distress (Matero et al., 2019).

In particular, research has shown the utility of LLMs in applications like depression detection from social media (Wolohan, 2020) and emotion recognition (Jiang et al., 2020). Other works have explored cross-lingual transformer-based psychological modeling (Yang et al., 2019) and the longitudinal study of personality traits through language (Eichstaedt et al., 2020). However, systematic comparisons of different LLM architectures (e.g., autoencoder, encoder-decoder, autoregressive) in their ability to model psychological constructs across varying levels of stability remain limited.

Our work expands on this foundation by systematically evaluating the capabilities of diverse LLM architectures to model psychological constructs. Unlike prior studies, we assess their performance across multiple temporal granularities (daily messages, bi-weekly waves, and aggregated user histories) using a longitudinal dataset. Furthermore, we investigate how different data aggregation strategies influence model performance, providing actionable insights for leveraging LLMs in psychological assessments.

3 Dataset

The dataset was collected over two years and comprises data from six waves, each lasting 14 days. Participants, U.S. restaurant workers, were recruited between June 2020 and June 2021 through service organizations and snowball sampling on social media. Enrollment was conducted via Qualtrics, and participants subsequently downloaded a custom smartphone app designed for ecological momentary assessment (EMA). Data collection began in 2021 with the first wave and con-

tinued with five additional waves throughout 2022.

To ensure the robustness of our analysis, only participants who contributed responses to at least two waves and provided a minimum of two responses per wave were included. This filtering resulted in a dataset containing 10,108 EMAs from 120 distinct users across 406 user waves.

Each day, participants reported the number of alcoholic drinks consumed in the past 24 hours and provided textual responses of at least 200 words to the prompt: *"Please describe in 2 to 3 sentences how you are currently feeling."* These responses were collected daily across all six waves.

Alongside daily responses, participants were asked once per wave to answer questions related to personality, mental health, affective states, stress, and alcohol abuse, self-reporting their scores for these parameters. The dataset thus contains both message-level data, comprising daily EMA responses such as textual descriptions of emotional states and alcohol consumption details, and wave-level data, which includes self-reported scores for personality, mental health, stress, affective states, and alcohol abuse. This longitudinal dataset provides valuable insights into the dynamic emotional and behavioral patterns of participants over time.

4 Methodology

This study systematically evaluates the effectiveness of 13 Transformer-based Language Models (LLMs) in capturing human psychological states and traits from textual data. We investigate three hierarchical levels of data analysis: **message-level**, focusing on individual Ecological Momentary Assessment (EMA) responses; **wave-level**, analyzing aggregated responses over 14-day periods; and **user-level**, aggregating data across all waves to capture long-term psychological trends.

Model Selection We evaluated a diverse range of Transformer-based models to capture psychological constructs, encompassing autoencoders: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), encoder-decoders: T5 (Raffel et al., 2020), FLAN-T5 (Chung et al., 2022), and autoregressive architectures: GPT-2 (Radford et al., 2019), HaRT (Soni et al., 2022), XLNet (Yang et al., 2019), Llama2 (Touvron et al., 2023b), Llama3 (Touvron et al., 2023a). Autoencoders provided dense contextual embeddings, encoder-decoders excelled at sequence-to-sequence tasks, and autoregressive models predicted tokens for coherent text generation. To assess the impact of architectural variations, we included both base and large variants of key models and fine-tuned instruction-based models like FLAN-T5. Specialized models such as HaRT, designed for human language modeling, leveraged hierarchical attention to capture granular textual patterns. Llama3-8B-Instruct was used for zero-shot prompting to evaluate performance without additional fine-tuning.

Outcomes In this study, we systematically analyzed a comprehensive range of psychological variables across multiple dimensions, including **affective** (Valence (Remmington et al., 2000), Arousal (Remmington et al., 2000), positive and negative affect: PANAS (Thompson, 2007)), **substance behaviors** (Number of Drinks, AUDIT-C, (Miller and Rollnick, 1993), Craving, MACE (Lange et al., 2017), **mental health** (depression: PHQ9 (Kroenke et al., 2003), anxiety: GAD7 (Plummer et al., 2016)), general stress: PSS (Cohen et al., 1983), daily stress: PSS Nervous Stress Agreement **socio-demographics** (Income, Age) and **personality** (BIG-5 (Soto, 2018): Openness, Neuroticism, Conscientiousness, Extraversion, Agreeableness). These variables were selected to capture a broad spectrum of psychological states and traits to evaluate the capability of open LLMs in modeling such constructs accurately and effectively.

All outcomes were self-reported by participants, with some collected at the EMA level and others at the Wave level. For Wave-level and User-level analyses of outcomes initially collected at the EMA level, the values were averaged across all EMA responses within each wave and subsequently across all waves for each user. Similarly, outcomes reported at the Wave level were averaged across all waves to derive User-level outcomes.

Stability To examine the stability of psychological constructs across different temporal levels, we computed two metrics: *intra-class correlation coefficients*(ICC) (Liljequist et al., 2019) and *test-retest correlations*(Pearson's r) (Weir, 2005). These metrics quantify the degree of consistency in self-reported outcomes over time, providing insights into the dynamic, moderately stable, or highly stable nature of each construct. Constructs with high stability exhibit minimal variability across EMAs or waves, while those with low stability are more dynamic and context-dependent.

Table 1 summarizes the stability metrics for both EMA-level and Wave-level analyses. Outcomes

Variables	EMA (~daily)		Wave (~quarterly)	
	ICC	retest	ICC	retest
Arousal (ARO)	0.108	0.173	0.443	0.490
Valence (VAL)	0.295	0.313	0.712	0.724
No of Drinks (DRI)	0.391	0.423	0.776	0.738
Craving (CRA)	0.412	0.481	0.607	0.726
Daily Stress(PSS1)	0.547	0.622	0.776	0.738
Stress(PSS)	-	-	0.580	0.586
Agreeableness (AGR)	-	-	0.638	0.617
Negative Affect (NAF)	-	-	0.602	0.642
Openness (OPE)	-	-	0.680	0.660
Conscientiousness (CON)	-	-	0.693	0.682
AUDIT C (AUC)	-	-	0.710	0.698
Positive Affect (PAF)	-	-	0.668	0.710
GAD7 (GAD)	-	-	0.720	0.736
PHQ9 (PHQ)	-	-	0.753	0.739
Neuroticism (NEU)	-	-	0.747	0.767
MACE (MAC)	-	-	0.775	0.775
Individual Income (INC)	-	-	0.768	0.793
Extraversion (EXT)	-	-	0.778	0.810
Age (AGE)	-	-	0.995	0.997

Table 1: Stability of the variables at both the EMA and wave Levels of analysis. Greater stability values indicate less change from EMA to EMA or wave to wave, respectively. ICC: intra-class coefficients; retest: average test-retest correlation (in Pearson r). Scores are highlighted to indicate standard categories: > 0.7: High (blue); 0.5 to 0.7: Medium (yellow); < 0.5: Low (green) (Koo and Li, 2016)

such as *Age* and *Personality Traits* (e.g., Extraversion, Conscientiousness) demonstrate high stability (ICCs and test-retest $r > 0.7$), as expected for constructs representing enduring traits. Moderately stable outcomes, such as *Stress* (PSS) and *Positive Affect*, have stability values in the range of 0.5 to 0.7. Dynamic constructs, including *Valence*, *Arousal*, and *Craving*, show lower stability (< 0.5), reflecting their sensitivity to momentary contextual changes.

These findings highlight the variability in construct stability, emphasizing the importance of tailoring modeling approaches to the temporal characteristics of each outcome. High-stability constructs benefit from aggregation strategies across waves or users, while low-stability constructs require models sensitive to fine-grained, momentary patterns.

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*.

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,
- r_{ij}: the Pearson correlation coefficient between temporal points i and j.

This approach ensures a robust measure of reliability by capturing the average consistency of variable measurements across all pairs of time points.

Intraclass Correlation Coefficient (ICC) 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.

These definitions are integral to selecting appropriate psychological measures and methodologies, ensuring that the data collected are reliable and valid for the intended research context.

Experimental Design In this study, we employed average token embeddings from selected models to predict specific psychological outcomes. To ensure a robust and unbiased evaluation, we utilized 10-fold cross-validation combined with ridge regression as our predictive modeling technique. Ridge regression was selected for its proficiency in addressing multicollinearity among predictors and its regularization capabilities, which help prevent overfitting. During cross-validation, data stratification was applied to maintain the distribution of outcome variables across all folds.

We explored two embedding aggregation strategies across three distinct levels of analysis: message, wave, and user. First, at the wave and user

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	0.57	35.99
RoBERTa-large	0.38**	3.04	0.37	8.07	0.39	30.85	0.59	33.71
BERT-base	0.35**	3.11	0.36	8.54	0.38	32.57	0.55	35.89
DeBERTa-base	0.37**	3.06	0.35	8.12	0.42	29.90	0.62	30.29
DeBERTa-large	**0.39****	3.00	**0.39**	8.27	0.41	30.70	0.62	30.79
Encoder-Decoder								
T5-large	0.37**	3.03	0.38	8.05	0.39	29.97	0.64	29.36
FLAN T5-large	0.38**	2.96	0.37	8.20	0.38	30.88	0.57	33.89
AutoRegressive								
GPT2-medium	0.36**	3.04	0.36	8.15	0.40	30.07	0.56	34.66
GPT2 HLC	0.36**	3.04	0.36	8.09	0.41	30.61	0.53	38.86
Xlnet-large	**0.39****	2.90	0.38	7.87	0.42	29.20	**0.68**	25.92
Llama2-7B	0.35*	3.04	0.36	8.52	0.37	31.97	0.54	35.18
Llama3-8B	0.34*	3.03	0.33	8.68	0.36	31.86	0.49	37.99
Human LM								
HaRT	0.34*	2.92	**0.39****	7.65	**0.44**	28.13	0.61	31.34
Zero Shot								
Llama3-8B	0.28	**4.48**	0.35	**73.87**	**0.44**	**55.83**	0.62	**52.86**

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.

levels, we concatenated all messages into single text sequences before generating their embeddings. This approach aimed to capture the cumulative context and interactions within each wave and across users, providing a comprehensive representation of psychological states over time or within individuals. Second, we generated embeddings for each individual message at the Ecological Momentary Assessment (EMA) level and then averaged these embeddings to create representative vectors at the wave or user levels. At the message level, embeddings were generated for each individual message to capture immediate linguistic and emotional cues. At the wave level, messages within the same time frame were aggregated for each user to create a more comprehensive embedding that reflects temporal patterns and fluctuations in psychological states. Finally, at the user level, embeddings were averaged across all messages and waves for each participant, providing a holistic representation of their overall psychological profile.

Evaluation Metrics. The performance of each model was evaluated using two primary metrics: Pearson correlation coefficient (r) and Mean Square Error (MSE). Pearson correlation measured the strength and direction of the linear relationship between the model predictions and self-reported scores, while MSE quantified the average magnitude of errors in the predictions compared to the labels.

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
GPT2 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	.280

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)

Computational Framework. The computational framework leveraged PyTorch, HuggingFace, and the Differential Language Analysis Toolkit (DLATK) (Schwartz et al., 2017) to extract meaningful features and evaluate model performance with precision. The framework was designed to handle the complexities of language-based data across varying levels of temporal granularity, ensuring robust and unbiased analyses. To support these operations, 2 NVIDIA RTX A6000 GPUs with 48GB of VRAM, were used for generating embeddings and executing zero-shot prompting tasks, enabling efficient processing and evaluation of the diverse LLM architectures used in this study.

5 Results

Capturing States, Dispositions, and Traits Table 2 presents the average Pearson correlation coefficients (r) that measure the accuracy of various LLMs in capturing three distinct categories of

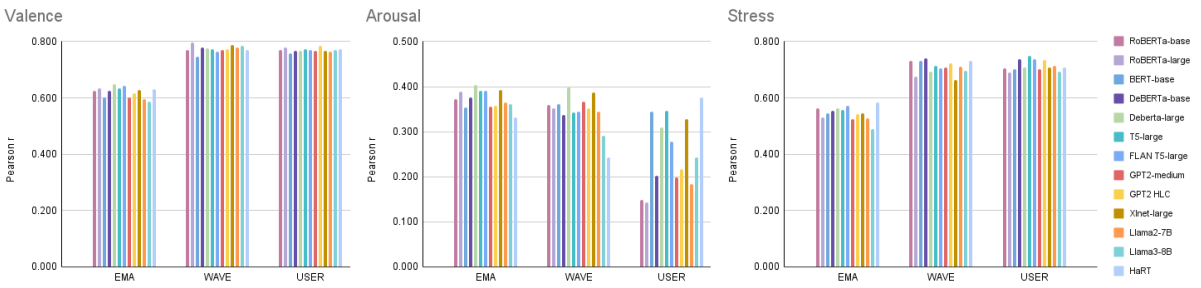

Figure 2: The predictive performance of different transformer-LM models across varying granularities—EMA, Wave, and User. Valence and stress are more accurately predicted at the wave or user level while arousal has greater accuracy at the EMA (i.e. document) level.

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

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 bootstrapped resampling across individuals over 1000 trials.

psychological constructs: states, dispositions, and traits. States are defined as variables with low stability, characterized by high variability across time, reflecting transient psychological conditions such as momentary emotions or acute stress responses. Dispositions exhibit moderate stability, representing semi-consistent psychological attributes that fluctuate to some extent, such as enduring moods or habitual behaviors. Traits are variables with high stability, indicating enduring and consistent psychological characteristics, such as core personality traits that remain relatively unchanged over extended periods. The table demonstrates the comparative performance of 13 Transformer-based LLMs in capturing psychological states, dispositions, and traits. Among autoencoders, DeBERTa-large consistently performed best across states (r = 0.39), dispositions (r = 0.39), and traits (r = 0.41), leveraging its advanced attention mechanisms to effectively model constructs of varying stability. In the encoder-decoder category, T5-large showcased balanced performance for traits (r = 0.39), while FLAN-T5-large excelled in capturing states (r = 0.38). Autoregressive models, particularly XLNet-large, demonstrated notable strengths in dynamic constructs, excelling in states (r = 0.39) due to its sequential modeling capabilities. The human-inspired HaRT model emerged as a leader for traits (r = 0.44) and dispositions (r = 0.39), highlighting the utility of hierarchical attention mechanisms for modeling stable constructs.

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. Results reveal distinct strengths among LLM architectures. HaRT, a Human LM that employs hierarchical attention mechanisms to model author context, consistently outperformed others, particularly excelling in affective constructs (r = 0.512), substance behavior (r = 0.321), mental health (r = 0.615), and personality (r = 0.342).

Predictive Performance Across Temporal Granularities Figure 2 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

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**	.442	.660	.495	.527	.052	.282	.231	.321	.418	.190	.426

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, while near-zero or slightly negative associations may arise due to differences in outcome prevalence between training and test samples, especially when sample sizes are small. Statistically significant differences from zero-shot Llama3-8B baseline: * ($p < .05$) and ** ($p < .001$), computed using boot-strapped resampling across individuals over 1000 trials. **Note**: See the Appendix Figure A1 for a more detailed discussion of these phenomena and additional experiments exploring model performance under varying data conditions.

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 As shown in Table 4, at the **EMA level**, reflecting immediate, state-like psychological constructs, DeBERTa-large achieved the highest correlations for emotional variables such as Valence ($r = 0.648$) and Arousal ($r = 0.404$), with FLAN-T5 closely following for Valence ($r = 0.642$). HaRT excelled in daily stress (PSS1, $r = 0.583$), number of drinks (DRI, $r = 0.296$), and Craving ($r = 0.317$), leveraging its hierarchical attention mechanisms, while XLNet-large showed strong performance in substance-related constructs such as DRI ($r = 0.271$) and Craving ($r = 0.281$). At the **wave level**, as shown in Table 5, which aggregates data over two-week periods to capture disposition-like patterns, embedding-based methods consistently outperformed zero-shot prompting across affective variables, mental health outcomes, personality traits, and demographic attributes. Autoencoder models, such as DeBERTa-large, effectively modeled moderately stable dispositions due to their advanced attention mechanisms, while autoregressive models like XLNet-large demonstrated strengths in specific dynamic constructs. Zero-shot prompting with models like Llama3-8B showed comparative advantages in substance-related behaviors, excelling in constructs such as craving and alcohol use. These findings emphasize the importance of matching model architecture to the stability and nature of psychological dimensions, with embeddings excelling in stable constructs and prompting better capturing dynamic, context-dependent behaviors. At the **user level**, as presented in Table 6, HaRT and zero-shot prompting methods, particularly Llama3-8B, performed best in capturing long-term, trait-like psychological constructs, including socio-demographics and stress-

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	.081	**.395**	.223	.406	.392	.135	.507

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, while near-zero or slightly negative associations may arise due to differences in outcome prevalence between training and test samples, especially when sample sizes are small. Statistically significant differences from zero-shot Llama3-8B baseline: * ($p < .05$) and ** ($p < .001$), computed using boot-strapped resampling across individuals over 1000 trials. **Note**: See the Appendix Figure A1 for a more detailed discussion of these phenomena and additional experiments exploring model performance under varying data conditions.

related measures. HaRT's hierarchical attention mechanisms made it particularly adept at stable constructs, while Llama3-8B showed strengths in substance-related behaviors, often surpassing traditional embedding-based methods. These results emphasize the importance of aligning model architecture and aggregation strategies with the stability and nature of psychological dimensions across temporal levels.

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 domains of measurement. 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 several limitations that stem from the ecological nature of its design. The participant population was restricted to bartenders and servers from professional and online groups within the United States, primarily individuals who may be at high risk for alcohol use due to their service industry roles. While participants were encouraged to provide responses three times a day over 14 days, some dropped out early or provided intermittent responses, resulting in incomplete time series that required interpolation techniques. Additionally, all participants were required to respond in English, and submissions in other languages or

spam-like responses were excluded. The sample also exhibited demographic skewness, with 75% of participants identifying as female, and the majority being middle-aged and located in the United States. This lack of diversity limits the generalizability of findings to broader populations and highlights the need for studies incorporating multilingual datasets and participants from varied demographic groups.

The small sample size further introduced computational challenges and restricted the scope of experiments that could be performed, particularly when evaluating hypotheses at a more granular level. Although the study assessed a broad range of psychological outcomes, it primarily focused on generalized disorders. Future research should expand these findings to more specific psychological conditions to enhance their applicability. Additionally, while this work evaluated a mix of smaller auto-encoder models and larger generative LLMs, expanding the analysis to include more state-of-the-art generative models would provide deeper insights into the differences between smaller and larger models in psychological modeling tasks, contributing significantly to the Computational Social Science (CSS) community.

8 Ethical Considerations

This study adheres to rigorous ethical guidelines to ensure the responsible application of artificial intelligence in mental health research. 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. The research protocol was reviewed and approved by an independent academic Institutional Review Board (IRB). 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

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. Technical report, University of Texas at Austin.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, et al. 2022. Scaling instruction-finetuned language models. arXiv 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.

Glen Coppersmith, Ryan Leary, Tony Whyne, and Benjamin Wood. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10:1–11.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 (NAACL-HLT)*, pages 4171–4186.

Johannes C Eichstaedt, Hansen Andrew Schwartz, and Margaret L Kern. 2020. The psychological language of social media: Associations between personality traits and linguistic patterns. *Nature Human Behaviour*, 4:980–991.

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

Fei Jiang, Ke Li, Jianwei Cui, and Hongying Zan. 2020. Cross-lingual emotion recognition with transformer-based models. arXiv preprint arXiv:xxxx.xxxxx.

Oscar Kjell et al. 2023. Transformer-based ai models approach theoretical upper-limit accuracy for psychological well-being assessments. *Psychological Methods*.

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.

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, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Matthew Matero, Daniel Idnani, Preethi Khattri, et al. 2019. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44.

Rohan Mehta, Sneha Kudugunta, Chandra Bhagavatula, and Tim Althoff. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4822–4832.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26.

William R. Miller and Stephen Rollnick. 1993. Motivational interviewing: Preparing people to change addictive behavior. *Addiction*, 88(6):849–869.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, et al. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

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.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, et al. 2015. Mental illness detection at the world well-being project for the clpsych 2015 shared task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pages 40–45.

Alec Radford, Jeffrey Wu, Rewon Child, et al. 2019. Language models are unsupervised multitask learners. OpenAI Blog. 1(8):9.

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.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2015. Using topic modeling to improve prediction of neuroticism and depression in college students. pages 1348–1353.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, et al. 2013a. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, 8(9):e73791.

H. Andrew Schwartz, Johannes C. Eichstaedt, Gregory Park, et al. 2013b. Characterizing geographic variation in well-being using tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 583–591.

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.

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 Soto. 2018. *Big Five personality traits*, pages 240–241.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

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, Peter Albert, Amjad Almahairi, Yasmine Babaei, Denis Bashlykov, Sharan Batra, Akshita Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Joseph T. Wolohan. 2020. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with bert. arXiv preprint arXiv:2004.11737.

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.

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|>
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|>
```

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 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

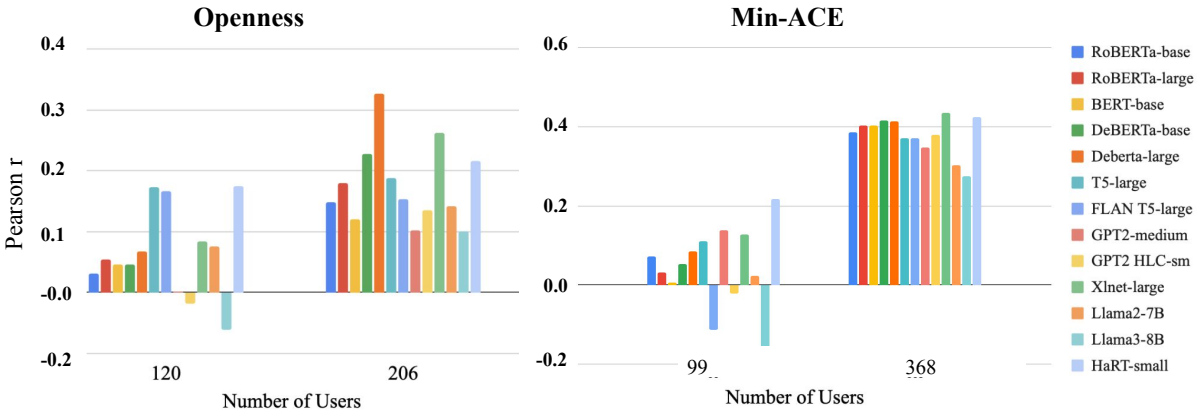

Figure A1: Model Performance vs Sample Size for Selected Outcomes. This figure illustrates how the predictive performance of different LLM-based embeddings (represented by distinct colored bars) varies as the number of users increases. For smaller sample sizes (shown on the left side of each plot), correlations 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.

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.
Anxiety	"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 score of the writer by analyzing their message. Anxiety score (ranging from 5 to 25) is calculated by assessing the presence and severity of indicators such as excessive worrying, agitation, restlessness, irritability, and physical symptoms like increased heart rate. Evaluate these factors based on explicit mentions, contextual descriptions, and implicit cues in tone and word choice. Each indicator is scored from 0 to 4, where 0 indicates absence and 4 indicates very high intensity or constant presence. Sum the individual scores and add 5 to align the result within the 5 to 25 scale, with 5 representing minimal anxiety and 25 indicating extremely severe anxiety. "
Depression	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 score of the writer by analyzing their message. Depression score ranging from 0 to 35 is calculated by evaluating the severity and frequency of depressive symptoms such as low mood, loss of interest in activities, reduced energy levels, negative self-perception, and sleep disturbances. Look for explicit mentions of these symptoms, descriptions of situations reflecting depressive states, and implicit cues in the tone, language, and overall mood conveyed in the text. Estimate the Depression score experienced by the writer, with 0 representing no depressive symptoms and 35 indicating very severe depressive symptoms.
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.
AUDIT10	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 40 using the AUDIT-10 scale, where 0 indicates no alcohol use or minimal risk, and 40 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 1 to 5, where 1 indicates very low levels of openness and 5 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 1 to 5, where 1 indicates very low levels of conscientiousness and 5 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 1 to 5, where 1 indicates very low levels of extraversion and 5 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 1 to 5, where 1 indicates very low levels of agreeableness and 5 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 1 to 5, where 1 indicates very low levels of neuroticism and 5 indicates extremely high levels of neuroticism .

Table A6: BIG-5 Personality Traits Variables