Evaluating Document-Tuned Transformer Representations for Person-level Mental Health Assessment

Aaron Marker
Vanderbilt University
aaron.marker@vanderbilt.edu

Oscar Kjell
Lund University

Vasudha Varadarajan
Carnegie Mellon University

H. Andrew Schwartz
Vanderbilt University
hansen.schwartz@vanderbilt.edu

Abstract

Person-level psychological assessment requires aggregating meaning across many messages from the same individual, a task that document-level training objectives were not explicitly designed for. We present a systematic, empirical comparison between architecturally matched traditional (a) *base-transformers* and (b) *document-tuned-transformers* (further contrastively fine-tuned at the document-level, sometimes referred to as "sentence transformers") under otherwise identical conditions. Comparing layer-wise and overall performance across two longitudinal mental health and psychological datasets, we find document-tuned models demonstrated a consistent improvement over base representations (increase in Pearson r of $13.4\%, p = .015$). Robustness analyses revealed document-tuned models remained more accurate under perturbations to word deletion, synonym replacement, typo injection, and back translation. Further, hedged language (e.g., 'usually') was more characteristic of outcomes in document-tuned embeddings while abundance (e.g., 'lot') was more characteristic of base-transformers, suggesting document-tuned models may better capture uncertainty. These results suggest representation choice impacts mental health prediction, document-tuned models often being more adept.

1 Introduction

Transformer-based language models are increasingly used to infer psychological constructs such as depression, anxiety, well-being, and personality from text (Tsakalidis et al., 2022; Zirikly et al., 2019; Guo et al., 2024; Pourkeyvan et al., 2024; Zhang et al., 2022). Much of prior work relies on token-level pretrained encoders (e.g., BERT, RoBERTa) (Liu et al., 2019b; Devlin et al., 2019). Recently, sentence transformers trained with document-level contrastive objectives (e.g., Sentence-BERT) have gained popularity for producing semantically meaningful sentence- and document-level embeddings (Reimers and Gurevych, 2019). These embeddings are typically frozen and used in a simple prediction head such that the quality of the representation itself becomes the primary driver of downstream performance.

Document-tuned transformers could bring two important advantages over the base transformers. First, psychological assessments are person-level outcomes (Zhang et al., 2022; Montejo-Raez et al., 2024; Garg, 2023) while base transformers ultimately represent information at a much lower level of aggregation, the token. Document-level representations already move to a higher level of aggregation (one per document or sentence) where by meaning at the level of whole sequences isn't relegated to words in context. Further, representations from traditional word-level transformers are also known to suffer from anisotropy (Ethayarajh, 2019), a geometric property where embeddings cluster in a narrow cone, rather than being uniformly distributed, often causing unrelated words to appear mathematically similar. Document-tuned transformers partially address both of these issues: tuned to produce document-level representations, they aggregate, and bring symmetry across dimensions of embeddings (Gao et al., 2021; Jung et al., 2023). However, empirically, it remains unclear which representation strategy is better suited for mental health assessment.

The psychological constructs targeted in this study span a broad range of clinical phenomena that differ in how they are experienced and expressed. Internalizing conditions such as depression and anxiety involves affective (e.g., nervous, sad), cognitive (e.g., useless, incapable), and somatic dimensions (e.g., headache, nausea) that individuals may disclose directly (Gu et al., 2025) or express through indirect, hedged language (Varadarajan et al., 2025). Substance use patterns, are often discussed obliquely — for example, problematic

Proceedings of the 11th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 178–187

drinking is associated with sadness, frustration, or even describing social events (Nilsson et al., 2024). Understanding how representation strategies may vary in their capacity to capture such diverse signals is essential for psychological prediction — particularly when prediction relies on aggregating many short, naturalistic text responses into a single person-level representation.

We conduct a controlled comparison of MLM-pretrained (base) and contrastively fine-tuned document-level (document-tuned) encoders for user-level psychological prediction. We aim to answer (RQ1) which model is more accurate for psychological prediction? (RQ2) what are optimal layer selection and message aggregation strategies? and (RQ3) do observed differences reflect sensitivity to specific linguistic features or perturbations? Our contributions are: (1) to our knowledge, the first controlled comparison of MLM-pretrained and document-tuned encoders for user-level psychological prediction; (2) an empirical analysis of layer selection and aggregation strategies across these encoder types; and (3) linguistic perturbation experiments that probe the source of observed performance differences in psychological settings.

2 Related Work

Much of prior work on transformer-based psychological prediction relies on token-level fine-tuned encoders. Alternative methods explicitly learn document-level representations optimized for semantic similarity. This distinction motivates a focused comparison of how representation choice impacts psychology-oriented prediction tasks.

2.1 Token level transformer fine-tuning

Pretrained transformer encoders, such as BERT, established a fine-tuning paradigm in which contextualized token-level models are fine-tuned end-to-end for downstream tasks (Devlin et al., 2019). Subsequent models, including RoBERTa, showed that architectural refinements and improved pretraining strategies can substantially improve performance, reinforcing token-level fine-tuning as a strong default across NLP tasks (Liu et al., 2019b). For sentence- or document-level prediction, these models typically rely on pooled token embeddings or a special classification token (e.g., [CLS]). However, because pretraining objectives are primarily token-level, these representations are not explicitly optimized for holistic document semantics.

2.2 Document level representation learning

A recent development focuses on learning general-purpose document embeddings. Sentence-BERT uses Siamese and contrastive objectives to produce fixed-size representations that capture semantic similarity (Reimers and Gurevych, 2019). Unlike token-level fine-tuning, these models directly encode entire sentences or short documents, enabling efficient reuse with lightweight downstream predictors.

2.3 Language modeling for psychological prediction

Transformer models have been widely applied in computational psychology, including tasks such as depression detection, suicide risk assessment, and mood prediction, often through shared tasks such as CLPsych (Tsakalidis et al., 2022; Tseriotou et al., 2025; Zhang et al., 2022; Montejo-Raez et al., 2024; Garg, 2023). While most work relies on token-level pretrained models, document-level encoders (e.g., Sentence-BERT) have also been used, for instance in ensemble systems and shared-task submissions (Ogunleye et al., 2024; Azim et al., 2022). The coexistence of these approaches highlights the need for controlled comparisons. We address this by comparing "roberta-large" and "all-roberta-large-v1" under matched extraction and evaluation settings.

3 Datasets

We evaluate our models on two longitudinal datasets that differ in scale and outcome focus: (1) a densely sampled ecological momentary assessment dataset (DS4UD) with broad psychological measures (Nilsson et al., 2024), and (2) a larger longitudinal mental health dataset (LMHD) with a broader user population (Kjell et al., in progress). Both datasets require aggregating multiple temporally distributed text responses to model person-level psychological outcomes, making them well-suited to examine how representation strategies handle aggregation across temporally distributed text.

3.1 The DS4UD Dataset

DS4UD is a longitudinal EMA dataset of U.S. service industry workers collected via a smartphone application (Nilsson et al., 2024). Participants (N=120; 75% female; $M_{age} = 35$) completed up to three EMAs per day across multiple 14-day waves.

Dataset	Statistic	
DS4UD	Total documents (EMAs)	10,108
	Average documents per person	84.9
	Total people	120
LMHD	Total documents (EMAs)	7,207
	Average documents per person	5.5
	Total people	1,307

Table 1: Descriptive statistics for the DS4UD and LMHD datasets under individual-message and concatenated-message representations.

Each EMA included an open-ended English text response (minimum 200 characters) describing the participant's current affective state, along with self-report ratings of momentary affect (positive and negative affect), perceived stress, recent alcohol use, alcohol craving, and energy.

At the start of each wave, participants completed standardized questionnaires assessing affect such as valence and arousal (Remington et al., 2000), positive and negative affect (Thompson, 2007), depressive symptoms (Patient Health Questionnaire–9; PHQ-9) (Kroenke et al., 2001), anxiety symptoms (Generalized Anxiety Disorder–7; GAD-7) (Spitzer et al., 2006), perceived stress (Perceived Stress Scale; PSS) (Cohen et al., 1983), alcohol use severity (Alcohol Use Disorders Identification Test; AUDIT, including the AUDIT-C) (Saunders et al., 1993), cravings, exposure to adverse childhood experiences (MACE) (Teicher and Parigger, 2015), and personality traits using a Big Five inventory (extraversion, agreeableness, conscientiousness, neuroticism, and openness) (Soto and John, 2017). These measures capture clinically relevant internalizing symptoms, substance use risk, stress, and stable personality characteristics. We retain participants who completed at least two waves, yielding over 10,000 text-based EMAs from 120 participants.

3.2 The LMH Dataset

The LMH Dataset comprise 10-week longitudinal data from a study evaluating alternative approaches to mental health assessment, including measures of general mental health and symptoms related to major depressive disorder (MDD) and generalized anxiety disorder (GAD). Participants were recruited via Prolific, and the present analyses include English-speaking participants only (N = 1,307; M age = 43.9 years, SD = 17.5; male = 34.3%, female = 64.0%, other = 1.7%). The sample was enriched: approximately 50% of participants were recruited from Prolific's general population, while the remaining participants were prescreened to report a prior diagnosis of MDD ($\approx$ 25%) or GAD ($\approx$ 25%).

Participants completed comprehensive surveys at baseline and follow-up, with shorter bi-weekly surveys administered in between. Across waves, the study assessed well-being; depressive and anxiety symptoms; perceived stress; alcohol and drug use; trauma-related symptoms; functional impairment; healthcare utilization; and sociodemographics using a combination of brief open-ended paragraph responses, word-selection formats, and standardized rating scales. Validated measures included the PHQ-9 and PHQ-2 for depressive symptoms; the GAD-7 and GAD-2 for anxiety ; subscales from the Inventory of Depression and Anxiety Symptoms (IDAS) assessing dysphoria, suicidality, panic, social anxiety, ill temper, lassitude, insomnia, appetite changes, traumatic intrusions, and well-being (Watson et al., 2007); the Perceived Stress Scale (PSS); the Alcohol Use Disorders Identification Test (AUDIT); the harmony in life score (HILS) (Kjell et al., 2016); the Drug Use Disorders Identification Test (DUDIT) (Berman et al., 2004); the Satisfaction with Life Scale (SWLS) (Diener et al., 1985); and the World Health Organization Disability Assessment Schedule (WHO-DAS) (Üstün, 2010). Additional items assessed sick days and mental health service utilization.

In the current study, we use the open-ended response to the general mental health prompt: "How is your mental health? Please describe how you have been over the last two weeks. You can, for example, write about your emotions, thoughts, behaviours, and/or symptoms related to your health. Write at least one paragraph."

4 Methods

We compare representations from (i) a masked language modeling (MLM) pretrained encoder (base) and (ii) a document-level contrastively tuned encoder (document-tuned) built on the same backbone. Representations are evaluated layer-wise under identical extraction, aggregation, and downstream prediction procedures across two longitudinal datasets.

4.1 Models

We compare `roberta-large` (Liu et al., 2019b), pretrained with masked language modeling as the base model, and `all-roberta-large-v1` variant as the document-tuned model, which fine-tunes the same backbone using just over one billion sentence-pairs for contrastive fine-tuning to produce fixed-size embeddings (Reimers and Gurevych, 2019; Hugging Face, 2025).

Both models share architecture and parameter count, isolating representation learning objective as the primary variable. We note that this comparison is not fully controlled for training data, as the document-tuned model was exposed to additional data beyond masked language modeling pretraining. We use publicly available HuggingFace implementations.

4.2 Representation extraction

We embed each message individually and mean-pool all message embeddings for a given user to obtain a fixed-length person-level representation. To assess whether document-tuned encoders benefit from longer context, we also evaluate a concatenation strategy in which all user messages are joined prior to embedding. When concatenated text exceeds the model context window, segments are embedded separately and mean-pooled.

4.3 Layer-wise representation selection

We extract and pool representations from every transformer layer of both encoders and train identical downstream models using nested cross-validation (described below). Layer-wise evaluation is motivated by prior work showing that linguistic information varies across depth (Tenney et al., 2019; Jawahar et al., 2019; Rogers et al., 2020; Liu et al., 2019a). Because document-tuned transformers are explicitly fine-tuned on a semantic similarity objective, they may contain more useful semantic signal in later layers relative to MLM-pretrained models. We therefore compare average Pearson's r score from the 10-fold cross validation, averaging across all outcomes in that dataset, across all layers and select the best-performing layer for each encoder.

4.4 Predictive modeling and cross-validation

For each outcome, we fit ridge regression models with nested 10-fold cross-validation, with the regularization parameter α selected via nested cross validation following prior work (Singh et al., 2025). Performance is measured using Pearson's r between predicted and true values, averaged across folds. The same downstream model and hyperparameter grid are used for all encoder–layer combinations.

5 Results and Discussion

5.1 Overall Accuracy and Optimal layers

Document-tuned representations outperform base model representations at nearly all layers across both datasets 2. Peak performance occurs in later layers for document-tuned transformers (Layer 21 on LMHD; Layer 19 on DS4UD) and earlier-to-mid layers for base models (Layer 19 on LMHD; Layer 10 on DS4UD). This pattern is consistent across datasets, suggesting document-level tuning objectives may preserve semantically relevant information in higher layers.

We compare models at their best-performing layers. Document-level representations achieve higher overall predictive performance on both datasets ($\Delta r = .012$ on LMHD; $\Delta r = .055$ on DS4UD; Table 6). Larger effect sizes in the DS4UD dataset could be due to the additional information of many more documents per person than in the LMHD dataset (DS4UD messages per person ≈ 85, LMHD messages per person ≈ 6)

5.2 Layer-wise Performance Analysis

Table 2 reports the top-performing layers (by Pearson's r) for both"roberta-large" (base) and "all-roberta-large-v1" (document-tuned) across datasets. This suggests that contrastive fine-tuning reorganizes semantic information toward the final layers, whereas MLM pretraining distributes it more evenly across depth — consistent with the view that later layers in document-tuned transformers are more specialized for semantic similarity.

5.3 Model Performance by Domain

Improvements are not uniform across outcome domains (Table 6). Gains are most pronounced for affect, personality, and especially drinking behavior measures ($\Delta r = .081$ on LMHD and .096 on DS4UD), and more modest or inconsistent gains for mental health and demographic outcomes. The only outcome where base models outperformed document-tuned models was for predicting a users craving, otherwise, differences were small and non-significant, consistent with chance variation around a null effect.

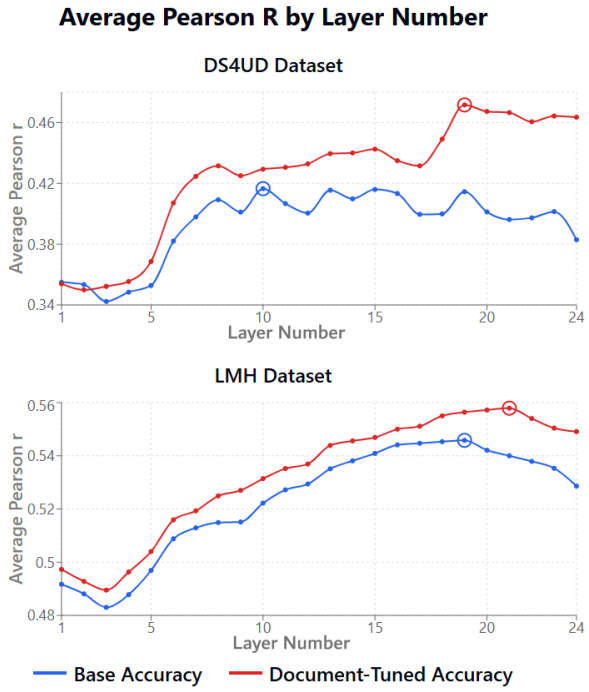

Figure 1: Prediction performance on the DS4UD (*N=120*) and LMHD (*N=1307*) datasets from different layer representations within "roberta-large" (Base Accuracy) and "all-roberta-large-v1" (Document-Tuned Accuracy)

Optimal Layer and Overall Performance								
	DS4UD (*N=120*)				**LMHD** (*N=1,307*)			
R	**L**	**Base**	**L**	**Doc**	**L**	**Base**	**L**	**Doc**
1	10	.4165	19	.4716	19	.5458	21	.5579
2	15	.4160	20	.4673	18	.5453 ↓	20	.5572
3	13	.4155	21	.4665	17	.5447	19	.5564 ↓
4	19	.4145	23	.4643 ↓	16	.5441 ↓	18	.5550 ↓
5	16	.4133	24	.4635 ↓	20	.5421 ↓	22	.5540 ↓
6	14	.4098	22	.4605 ↓	15	.5409 ↓	17	.5511
7	8	.4092	18	.4491 ↓	21	.5400 ↓	23	.5504 ↓
8	11	.4067	15	.4425 ↓	14	.5381 ↓	16	.5500
9	23	.4014	14	.4400 ↓	22	.5379 ↓	24	.5491 ↓
10	20	.4012	13	.4395 ↓	23	.5353 ↓	15	.5469 ↓

Table 2: Top-performing (Pearson r) layers for the DS4UD and LMHD dataset for "roberta-large" (Base) and "all-roberta-large-v1" (Doc). ↓ indicates significant distinction from the best performing layer ($p < .05$).

Model Performance				
Outcome Category	**Dataset**	**Doc**	**Base**	**Δ**
Overall	LMHD	0.572	0.561	.012*
	DS4UD	0.472	0.416	.055*

Table 3: Document-tuned vs. base model performance (Pearson r) by dataset. Δ = document-tuned − base. $^{*}p < .05$, $^{**}p < .01$.

Concatenating all user messages prior to embedding only slightly increases the advantage of document-tuned transformer embeddings, but relative to mean pooling, results are inconsistent, and on the LMHD dataset, drops both embedding types' predictive accuracy (Appendix table 4).

5.4 Supervised Dimension Projection

To further contextualize these findings, we qualitatively examine Supervised Dimension Projection Plots of specific words. These plots are constructed by computing an embedding vector representing the difference between high and low values of a given outcome. Individual words are then projected onto this axis using dot product projections, indicating how strongly each word is associated with the outcome in a given embedding space (Kjell et al., 2023). To compare two models' embeddings of the same dataset, we plot the word along an embedding dimension representing the outcome in both embedding spaces, resulting in a 2D graph. We find no consistent trends in the number of significantly predictive words for either model. While graphs show both expected behavior (like "happy" being equally low in both representations), it also shows words like "Alright" in the second quadrant, which maps to high PHQ scores in document-tuned embeddings and low in base embeddings. This suggests that document-tuned models may encode the word 'Alright' in a context more associated with 'not being alright', whereas base models may place more emphasis independently on its definition alone — potentially reflecting the richer contextual representations learned through contrastive training. Embedded hedged language (e.g., 'usually' in PHQ and 'pretty' in AUDIT) tended to be more predictive of outcomes in document-tuned embeddings while abundance (e.g., 'lot') was more predictive in traditional word-level models, suggesting document-tuned models may better capture uncertainty and tentativeness in self-report language.

5.5 Perturbations

In recent work, such as (Alahmari et al., 2025), models have been probed for their robustness to human-induced linguistic errors that are common in real-world applications. As natural language surveys and EMAs are prone to user error, we corrupt text using deletion, typo injection, synonym replacement, and back translation at varying levels to simulate increasingly noisy data. We find

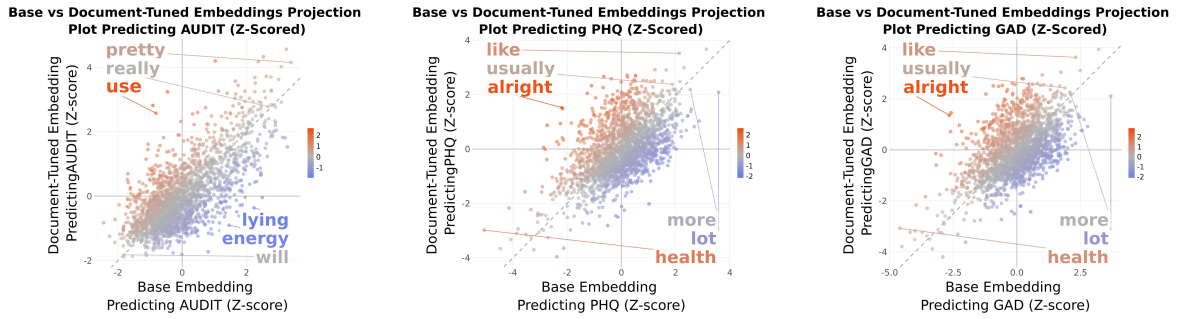

Figure 2: Dimension Projection Plot where words from the LMHD dataset are plotted along the embedding vector representing a difference in High vs Low PHQ. Words plotted along the diagonal line are represented similarly in both embedding spaces with relation to PHQ. More saturated points (like "meaning") are associated with high PHQ in one space and low PHQ in the other. Orange points are more highly associated with the outcome in Document Tuned embeddings, and blue points are more highly associated with the outcome in Base embeddings.

Model Performance by Message Combination Method				
Dataset	**Message Comb.**	**Doc**	**Base**	**Δ**
LMHD	Mean Pool	0.572	0.561	.012*
	Concatenate	0.546	0.532	.014*
DS4UD	Mean Pool	0.472	0.416	.055*
	Concatenate	0.482	0.411	.071**

Table 4: Document-tuned vs. base performance (Pearson r) by dataset and message combination. Δ = document-tuned − base. $^{*}p < .05$, $^{**}p < .01$.

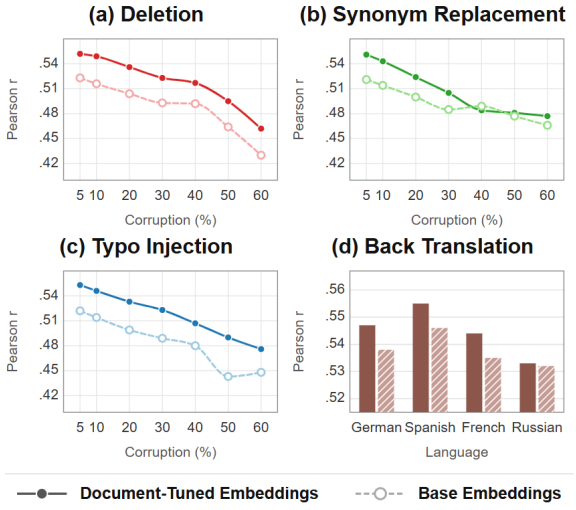

Figure 3: Varying Levels of Text Perturbations vs Model Prediction Correlation with True Scores. (a) consists of deleting varying percentages of words. (b) consists of adding, replacing, or reordering adjacent letters in varying percentages of words. (c) consists of using WordNet to replace varying percentages of words with synonyms. (d) consists of translating the message into another language and back again.

in most cases, neither document-tuned nor base transformers are more robust to these perturbations than the other. Across deletion, typo injection, and back translation, both models showed comparable and gradual performance degradation as corruption levels increased. Synonym replacement shows slightly more robustness by the base transformer at extremely high percentages of replaced words, potentially because base model embeddings are less sensitive to global semantic context, making synonym substitution less disruptive to the overall representation. This fails to explain the stronger overall performance of document-tuned models, suggesting that their advantage stems from a more fundamental difference in how meaning is encoded that does not simply make the model more robust to syntactic or semantic noise.

6 Conclusion

This work presents a controlled comparison of base transformers (utilizing token representations) and document-tuned transformers (utilizing document-level, contrastively tuned representations) for psychological prediction. Document-tuned transformers consistently outperform base transformers for user-level psychological prediction across two longitudinal datasets. Gains are modest, but statistically significant and robust across layers. Optimal performance occurs in later layers for document-tuned models, with the final four to five layers performing comparably. We examined several candidate explanations for the observed performance differences, including outcome type, word embedding representation, and robustness to linguistic perturbations, but found no consistent moderating patterns. Crucially, the advantage of document-tuned representations was not confined to any par-

ticular outcome domain, suggesting a broadly superior capacity to encode person-level meaning rather than sensitivity to specific linguistic features — making document-tuned models a reasonable default choice for language-based mental health assessment. For researchers developing language-based screening or monitoring systems – particularly those using EMA or open-ended survey responses – these results provide empirical guidance on a design decision that is typically made without systematic comparison: which encoder to use and which layers to extract embeddings from.

7 Ethical Considerations

Both datasets were collected under institutional review board (IRB) approval with informed consent from participants. Analyses were conducted on de-identified text and survey data.

In particular, the improved prediction of sensitive outcomes such as drinking behavior and suicidality from open-ended text raises questions about the potential for unintended inference in deployed systems. Models developed in this work should not be used for clinical diagnosis or high-stakes decision-making without appropriate validation and oversight. While our study focuses on methodological comparison rather than deployment, representation choices that improve predictive performance may also increase risks of unintended inference. Care should be taken to ensure that such systems are used transparently, with appropriate safeguards, and in ways that respect participant privacy and autonomy.

8 Limitations

Several limitations should be noted. First, our comparison uses off-the-shelf checkpoints. "All-roberta-large-v1" includes additional post-pretraining beyond masked language modeling, involving different data. As such, observed gains cannot be attributed solely to the contrastive objective without compute- and data-matched controls.

Second, we evaluate only two English-language longitudinal datasets, including one relatively small sample (N=120). Results may not generalize to other populations, languages, platforms (e.g., clinical notes or social media), or measurement settings.

Third, our evaluation focuses on predictive accuracy (Pearson's r) and does not assess calibration, fairness, or error disparities across demographic groups, which are important considerations for assessment contexts.

Fourth, though this dataset contains longitudinal data, we chose to measure cross sectional prediction accuracy for simplicity and ease of comparison between the datasets. We leave studying longitudinal models of psychological constructs to future work.

Finally, this study evaluates methodological differences rather than real-world deployment. Improvements in predictive performance do not imply clinical validity or suitability for high-stakes decision-making.

References

Saeed S Alahmari, Lawrence Hall, Peter R Mouton, and Dmitry Goldgof. 2025. Large language models robustness against perturbation: S. alahmari et al. *Scientific Reports*.

Tayyaba Azim, Loitongbam Gyanendro Singh, and Stuart E Middleton. 2022. Detecting moments of change and suicidal risks in longitudinal user texts using multi-task learning. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 213–218.

Anne H Berman, Hans Bergman, Tom Palmstierna, and Frans Schlyter. 2004. Evaluation of the drug use disorders identification test (dudit) in criminal justice and detoxification settings and in a swedish population sample. *European addiction research*, 11(1):22–31.

Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior*, pages 385–396.

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, volume 1 (long and short papers)*, pages 4171–4186.

ED Diener, Robert A Emmons, Randy J Larsen, and Sharon Griffin. 1985. The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Muskan Garg. 2023. Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering*, 30(3):1819.

Zhuojun Gu, Katarina Kjell, H Andrew Schwartz, and Oscar Kjell. 2025. Natural language response formats for assessing depression and worry with large language models: A sequential evaluation with model pre-registration. *Assessment*, page 10731911251364022.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, and 1 others. 2024. Large language models for mental health applications: systematic review. *JMIR mental health*, 11(1):e57400.

Hugging Face. 2025. sentence-transformers/all-roberta-large-v1. https://huggingface.co/sentence-transformers/all-roberta-large-v1. Accessed: 2026-01-??

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Euna Jung, Jungwon Park, Jaekeol Choi, Sungyoon Kim, and Wonjong Rhee. 2023. Isotropic representation can improve dense retrieval. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 125–137. Springer.

Oscar Kjell, Salvatore Giorgi, and H Andrew Schwartz. 2023. The text-package: An r-package for analyzing and visualizing human language using natural language processing and transformers. *Psychological methods*, 28(6):1478.

Oscar NE Kjell, D Daukantaitė, Kate Hefferon, and Sverker Sikström. 2016. The harmony in life scale complements the satisfaction with life scale: Expanding the conceptualization of the cognitive component of subjective well-being. *Social Indicators Research*, 126(2):893–919.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

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

Arturo Montejo-Raez, M Dolores Molina-Gonzalez, Salud Maria Jimenez-Zafra, Miguel Angel Garcia-Cumbreras, and Luis Joaquin Garcia-Lopez. 2024. A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Computer Science Review*, 53:100654.

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.

Bayode Ogunleye, Hemlata Sharma, and Olamilekan Shobayo. 2024. Sentiment informed sentence bert-ensemble algorithm for depression detection. *Big Data and Cognitive Computing*, 8(9):112.

Alireza Pourkeyvan, Ramin Safa, and Ali Sorourkhah. 2024. Harnessing the power of hugging face transformers for predicting mental health disorders in social networks. *IEEE Access*, 12:28025–28035.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nancy A Remington, Leandre R Fabrigar, and Penny S Visser. 2000. Reexamining the circumplex model of affect. *Journal of personality and social psychology*, 79(2):286.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

John B Saunders, Olaf G Aasland, Thomas F Babor, Juan R De la Fuente, and Marcus Grant. 1993. Development of the alcohol use disorders identification test (audit): Who collaborative project on early detection of persons with harmful alcohol consumption-ii. *Addiction*, 88(6):791–804.

Khushboo Singh, Vasudha Varadarajan, Adithya V Ganesan, August Håkan Nilsson, Nikita Soni, Syeda Mahwish, Pranav Chitale, Ryan L Boyd, Lyle Ungar, Richard N Rosenthal, and 1 others. 2025. Systematic evaluation of auto-encoding and large language model representations for capturing author states and traits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18955–18973.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.

Martin H Teicher and Angelika Parigger. 2015. The 'maltreatment and abuse chronology of exposure'(mace) scale for the retrospective assessment of abuse and neglect during development. *PLoS one*, 10(2):e0117423.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

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.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, and 1 others. 2022. Overview of the clpsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198.

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, and 1 others. 2025. Overview of the clpsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 193–217.

T Bedirhan Üstün. 2010. *Measuring health and disability: Manual for WHO disability assessment schedule WHODAS 2.0*. World Health Organization.

Vasudha Varadarajan, Allison Lahnala, Sujeeth Vankudari, Akshay Raghavan, Scott Feltman, Syeda Mahwish, Camilo Ruggero, Roman Kotov, and H Andrew Schwartz. 2025. Linking language-based distortion detection to mental health outcomes. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 62–68.

David Watson, Michael W O'Hara, Leonard J Simms, Roman Kotov, Michael Chmielewski, Elizabeth A McDade-Montez, Wakiza Gamez, and Scott Stuart. 2007. Development and validation of the inventory of depression and anxiety symptoms (idas). *Psychological assessment*, 19(3):253.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

A Appendix

A.1 Full Data Sizes

Table 5 summarizes data from both datasets.

		EMAs	
Dataset	**Statistic**	**Indiv.**	**Concat.**
DS4UD	Avg. EMA per emb.	1	6.9
	Avg. emb. per user	84.9	12.4
	Total embeddings	10,108	1,486
	Total users	120	120
LMHD	Avg. EMA per emb.	1	5.2
	Avg. emb. per user	5.5	1.1
	Total embeddings	7,207	1,391
	Total users	1,307	1,307

Table 5: Descriptive statistics for the DS4UD and LMHD datasets under individual-message and concatenated-message representations.

A.2 Outcome-Level Results by Domain

Table 6 summarizes performance by outcome category.

Model Performance by Outcome Category				
Outcome Category	**Dataset**	**Doc**	**Base**	Δ
Affect	DS4UD	0.502	0.432	.070*
Demographic	DS4UD	0.489	0.429	.061
Drinking behavior	LMHD	0.263	0.182	.081*
Drinking behavior	DS4UD	0.337	0.241	.096**
Mental health	LMHD	0.576	0.566	.010*
Mental health	DS4UD	0.636	0.615	.021
Personality	DS4UD	0.378	0.337	.041*
Overall	LMHD	0.572	0.561	.012*
	DS4UD	0.472	0.416	.055*

Table 6: Document-Tuned vs. Base model performance (Pearson r) by outcome category and dataset. Δ = Document-Tuned $-$ Base $^{*}p < .05$, $^{**}p < .01$.

A.3 Full Outcome-Level Results

Table 7 reports prediction performance for all individual outcomes across both datasets.

Model Performance							
Outcome	**N**	**MAE**			**Pearson r**		
		Doc	**Base**	**Δ**	**Doc**	**Base**	**Δ**
LMHD							
AUDIT	1307	3.63	3.70	$-.07^{*}$	.263	.182	$.081^{*}$
Age	1307	8.62	8.14	.48	.792	.812	$-.020$
DUDIT	1307	1.79	1.78	.01	.181	.156	.025
GAD2	1307	1.12	1.10	.02	.662	.673	$-.011$
GAD	1307	3.05	3.09	$-.04$	.724	.725	$-.001$
HILS	1307	2.87	2.88	$-.01$	.710	.705	.005
HCMental	708	0.83	0.83	.00	.339	.349	$-.011$
HCVisits	1307	1.34	1.33	.01	.357	.321	.036
IDAS_AppGain	1307	2.63	2.64	$-.01$	.300	.291	.009
IDAS_AppLoss	1307	2.08	2.08	.00	.412	.415	$-.004$
IDAS_Dysphoria	1307	5.05	5.04	.01	.752	.753	$-.001$
IDAS_Temper	1307	2.41	2.47	$-.06^{*}$	.573	.542	$.031^{*}$
IDAS_Insomnia	1307	4.55	4.59	$-.04$	.475	.459	.016
IDAS_Lassitude	1307	3.14	3.09	.05	.676	.686	$-.009$
IDAS_Panic	1307	3.24	3.35	$-.11^{*}$	.585	.562	$.022^{*}$
IDAS_SocAnx	1307	3.15	3.13	.02	.610	.614	$-.005$
IDAS_Suic	1307	1.94	1.97	$-.03$	.604	.576	.028
IDAS_Traum	1307	2.31	2.36	$-.05^{*}$	.576	.545	$.031^{*}$
IDAS_WB	1307	4.87	5.01	$-.14^{*}$	.670	.665	$.005^{*}$
MentalSickDays	395	12.56	12.64	$-.08$	.390	.354	.036
PHQ2_sum	1307	0.89	0.90	$-.01$	.712	.705	.007
PHQtot	1307	3.29	3.32	$-.03$	.744	.739	.005
PSS4_sum	1263	2.11	2.11	.00	.692	.688	.004
PSStot	1307	4.87	4.89	$-.02$	.741	.740	.001
SWLStot	1307	3.10	3.14	$-.04$	.683	.676	.007
SickDaysMonth	1307	5.63	5.56	$.07^{*}$	.206	.189	$.018^{*}$
WHODAS	1307	17.47	17.96	$-.49^{*}$	.636	.615	$.021^{*}$
Mean		4.02	4.04	$-.02^{*}$	.572	.561	$.012^{*}$
DS4UD							
affect	120	0.35	0.39	$-.04^{*}$	.799	.752	$.047^{*}$
age	103	5.71	5.95	$-.24$	.500	.425	.074
agreeable	120	1.75	1.76	$-.01$	.455	.393	.063
anxiety	120	2.65	2.70	$-.05$	.582	.568	.014
audit10	103	5.41	5.70	$-.29^{*}$	.318	.130	$.188^{*}$
auditc	120	1.94	2.07	$-.13^{*}$	.310	.214	$.097^{*}$
conscientious	120	2.01	2.10	$-.09^{*}$	.392	.372	$.020^{*}$
craving 1	94	1.21	1.19	.02	.403	.439	$-.036$
depression score	120	3.90	3.81	.09	.600	.601	$-.002$
energy	120	0.24	0.25	$-.01^{*}$	.283	.129	$.154^{*}$
extravert	120	2.64	2.63	.01	.320	.304	.017
gad7 sum	120	3.64	3.65	$-.01$	.610	.567	.042
income	120	2.00	2.01	$-.01$	.479	.432	.048
mace	99	9.26	10.27	-1.01^{*}	.339	.160	$.179^{*}$
neg affect	103	3.03	3.12	$-.09$	.536	.510	.026
neurotic	120	2.00	2.10	$-.10$	.632	.586	.046
openness	120	2.23	2.23	.00	.089	.030	.059
phq9	120	3.72	3.77	$-.05$	.601	.594	.007
pos affect	103	3.21	3.38	$-.17^{*}$	.389	.335	$.054^{*}$
pss nerv. stress agr.	92	0.57	0.62	$-.05^{*}$	.784	.722	$.061^{*}$
pss	120	1.66	1.67	$-.01$	.639	.635	.004
unhealthy drinking	120	1.28	1.31	$-.03$	.314	.264	.050
Mean		2.75	2.85	$-.10^{*}$	.472	.416	$.055^{*}$

Table 7: Document-tuned vs. base performance by outcome and dataset. MAE = Mean Absolute Error (lower is better); Pearson r (higher is better). Δ = Document-tuned $-$ Base. $^{*}p < .05$, $^{**}p < .01$ (significances from t-test with MAE).