Text and Valence-Arousal: A Two-Dimensional Foundational Approach for Mental Health Prediction

Christo Hristov, Ion Martinis, Yalcin Tur, Kevina Wang, Miko Rimer
[christoh], [ion2004], [yalcintr], [kevinaw], [mikor] @ stanford.edu
Stanford University – CS 277 / BIODS 271
Spring 2025

Abstract

We present a two-dimensional foundational framework for mental health assessment that leverages continuous valence-arousal (VA) coordinates extracted from natural language as a universal affective marker. Rather than training disorder-specific models, we developed a generalizable VA regressor using EmoBank and GoEmotions datasets, with our optimal RoBERTa architecture achieving a 0.0759 MAE and 0.82 valence correlation. We later demonstrate cross-disorder applicability by augmenting clinical transcripts with VA coordinates for both depression (PHQ-8) and PTSD (PCL-5) prediction tasks on the E-DAIC dataset. VA-enhanced prompts yield consistent improvements across model paradigms: 12.2% MAE reduction in supervised fine-tuning for depression, and 10.7% accuracy gains for PTSD classification. Crucially, our temporal trajectory analysis reveals that depressed individuals exhibit characteristic negative valence excursions and increased emotional volatility throughout clinical interviews, while PTSD-positive participants show heightened arousal variability. Unlike existing approaches that target single conditions, our VA framework provides a foundational affective substrate that is generalizable across mental health domains, enabling real-time emotional monitoring through passive language analysis without requiring disorder-specific fine-tuning.

1 Introduction

Nearly 1 billion people worldwide suffer from mental disorders (Global Burden of Disease 2017 Collaborators, 2018). This immense global burden – worsened by recent crises and a shortage of mental health providers (Garriga et al., 2022) – underscores the urgent need for improved diagnostic and monitoring tools (Garriga et al., 2022). However, clinicians still rely heavily on self-administered questionnaires, which give only static snapshots of a patient's condition and lack temporal sensitivity to fluctuations in mental state (Zimmerman, 2024). Although these approaches have, from an epidemiological level, enabled clinicians to inquire about certain disorders on which the patient screens positive, they are limited in their utility for timely intervention as they are self-administered (Zimmerman, 2024).

Natural language remains a principal component of mental health assessment (Wang et al., 2024), and as a result recent advances in machine learning have increased the ability to develop and deploy digital mental health models. However, current approaches have notable limitations, as many models are domain specific, resulting in models that are narrowly tuned to particular diagnoses (Ji et al., 2021) (Wang et al., 2024). This limits their generalizability across the diverse spectrum of mental health conditions, as they are pretrained on specific disorders (Wang et al., 2024). Similarly, computational emotion models, such as GPT-4, identify situational context by computing how specific textual components may evoke particular emotions – for example, classifying the emotion *grateful* from text (Tak and Gratch, 2024). Such models often ignore the established view that emotions vary along continuous dimensions of valence (sad to cheerful) and arousal (quiet to active) (Bradley and Lang, 1994) (Jia et al., 2024). Moreover, modeling emotion using discrete dictionaries introduces inherent bias, as it represents mental state through classifications of specific emotions selected for their association with symptom pathology, rather than capturing affect on a standardized, continuous scale across a diverse range of emotional states (Jia et al., 2024). Although certain models like Garriga et al. (2022)'s algorithm reduced crisis risk in 64% of cases, they rely on structured electronic health record data and frequent patient contact, making them impractical outside specialized settings where continuous clinical records are available. In summary, existing solutions either lack real-time sensitivity, broad applicability, or the ability to represent emotional states holistically.

To address this gap, we propose a novel approach to use a foundational emotional metric derived from a text input that can serve as a proxy for mental state. We introduce a language-based metric of emotional state – defined by valence and arousal (VA) – as a compact, interpretable two-dimensional representation of mental health. From a broader perspective, our model can infer a range of continuous metrics from any form of natural language, offering a passive and scalable approach to real-time monitoring of affective states. This foundation supports proactive mental health care by continuously tracking emotional trajectories, moving the field beyond episodic check-ins toward early digital-detection and prevention in a wide range of mental health conditions.

2 Related Work

Most existing text-based mental health models are task-specific, built to detect one single-granular issue (e.g., depression or suicide risk) in a single data domain (such as therapy transcripts or Reddit forums). For example, Sadeghi et al. (2024)'s convolutional neural network detects early depression symptoms from text, audio and video input, using a fine-tuned RoBERTa (Pourkeyvan et al., 2024) transformer models (DepRoBERTa) to perform feature extraction of depression from text. Similarly, Pourkeyvan et al. (2024)'s "MentalBERT was developed based on posts containing mental health-related information collected from Reddit and is based on BERT-Base". Current evaluation benchmarks for mental health prediction are limited; however, the Extended DAIC-WOZ (E-DAIC) corpus serves as a relevant silver-standard reference (Gratch et al., 2014) (Ringeval et al., 2019). It has been noted that that E-DAIC may overestimate performance due to the presence of gender bias and interviewer prompts that inadvertently simplify assessments (Burdisso et al., 2024). Nevertheless, the E-DAIC dataset does offer a well—structured benchmark widely utilized across numerous studies over time (Sadeghi et al., 2024). Previous approaches have primarily leveraged this dataset to predict depression severity using discrete labels based on the PHQ-8 (ranging from 0 to 24). Because these systems are fine-tuned for a single context and provide only discrete scores, they fail to capture deeper insights into an individual's underlying emotional state beyond the depression prediction. Recent work has attempted to enhance PHQ-8 prediction accuracy by explicitly integrating topic-wise emotional tendencies, like Guo et al. (2024)'s method to predict emotional trajectories of each interview segment with emotionally polarized prompts, think postivie or negative. To our knowledge, there is yet to be a model that can broadly capture mental state across multiple benchmarks (think a unified model to capture PHQ-8 and PTSD severity from the E-DAIC).

Our work addresses these gaps by learning a generalizable affective representation as a foundation. We first train our model within a two-dimensional VA space using similar methods from Mitsios et al. (2024) and Jia et al. (2024)'s approaches by using large-scale emotion datasets (EmoBank and GoEmotions). The result is an initial VA prediction model that is explicitly in a foundational dimensions that captures human affect. We then explored multiple approaches for incorporating this VA context to enhance the detection of PHQ-8 scores, and subsequently PTSD symptoms, from the E-DAIC dataset, measuring improvements using mean absolute error (MAE).

3 Approach

Our approach was in two parts: **1)** building a regressor model that can infer VA from text, **2)** evaluating the utility of our VA regressor for tracking mental state by inputting VA-labelled clinical transcripts into promptable foundation models and predicting scores of depression and PTSD.

3.1 Valence-Arousal Prediction

We first developed a regressor model that predicts valence-arousal (VA) associated with text. We evaluated three approaches: 1) GPT o4-mini with zero-shot prompting, 2) RoBERTa with emotional context conditioning, and 3) RoBERTa for direct VA regression.

3.1.1 Base GPT

This involved GPT-o4-mini with zero-shot prompting, the task being to predict VA scores from text. Since we planned to input text-VA pairs to GPT models for the second part of our approach, we wanted to evaluate if GPT could be used end-to-end for both VA and mental health score prediction.

3.1.2 RoBERTa Direct for VA Regression

We sought to train a VA regressor model based on a large pre-trained language transformer. We selected RoBERTa as the pre-trained model due to it being encoder-only, lighter in model parameters, and open-source allowing us directly optimize for the VA task.

We considered a range of architectures with fully-connected layers stacked on top of the encoder and a final tanh to project to VA in the range [-1, 1]. The best performing architecture is in Figure 1 (below).

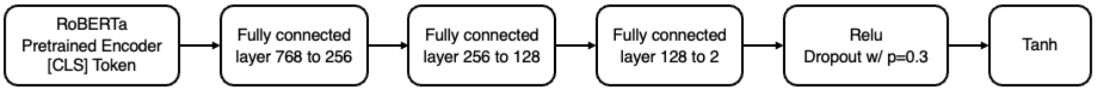

Figure 1: Direct VA Regression Model

3.1.3 RoBERTa with Emotion Context

Given that VA describes emotional state, we hypothesized that initially fine-tuning a large transformer model to understand emotion classes could enhance the model's ability to learn the VA prediction task downstream. For our emotion-aware transformer, we used the model presented by Kashyap (2024) which consists of a RoBERTa encoder fine-tuned on the GoEmotions dataset (mapping of text to 28 discrete emotion labels). We repeated the authors' training procedure and achieved similar performance of 0.74 macro F-1 score. The model architecture is shown in Figure 2 (below).

Next, we added fully connected layers and a final projection layer on top of the $768 \rightarrow 256$ layer trained on GoEmotions (Figure 3, below). Initializing with the model weights from GoEmotions training, we trained this extended architecture for the VA prediction task. To initially preserve the emotion context from training on GoEmotions, we started with a frozen RoBERTa encoder and $768 \rightarrow 256$ layer, which we unfroze after validation performance had plateaued to push to an optimum.

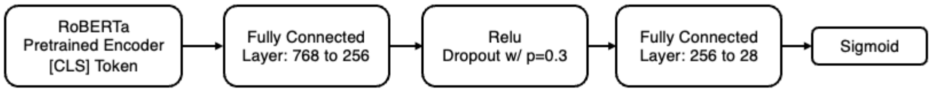

Figure 2: RoBERTa model fine-tuned on GoEmotions as presented by Kashyap (2024)

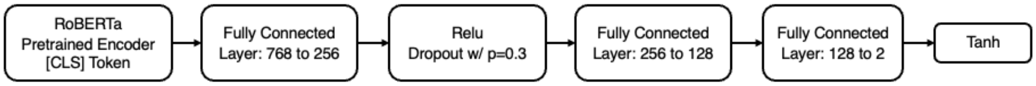

Figure 3: Extended RoBERTa model for VA regression task

3.2 Prediction of depression and PTSD scores from text + VA as context

For this part, we evaluated the utility of VA for mental health assessment. Using our best model for VA prediction, we performed inference of VA for each line of psychiatric clinical transcripts, then we inputted the transcripts with their VA labels to GPT models to predict scores of PHQ-8 (depression) and PCL-5 (PTSD) separately. The prompt for the GPT models stated the prediction task and included the full clinical transcript labelled with inferred VA. We evaluated three GPT-based approaches for predicting PHQ-8 and PCL-5 scores.

3.2.1 Base GPT with zero-shot prompting

This baseline approach used GPT-4o-mini without training examples, with structured prompts instructing the model to predict mental health scores from VA-annotated transcripts. This tests whether foundation models can inherently utilize continuous affective signals for clinical assessment.

3.2.2 AutoCoT

We hypothesized that by providing GPT few shots of VA-labelled transcripts and ground-truth PHQ-8/PCL-5 scores, the model could learn in-context how VA might map to mental health assessment. As few shots, we used transcripts from the E-DAIC training set. Following the AutoCoT framework, we passed few shots to a separate GPT-o4-mini prompted with the following format:

```
Q: [Reasoning instruction]
A: [Zero-shot Generated rationale]. Therefore, the answer is
[Ground truth PHQ-8].
```

Our goal with AutoCoT was to generate rationales of few shots and guide the model to perform structured reasoning. At inference time, we chained formatted demos of few shots with patient transcripts and tasked the model to predict PHQ-8/PCL-5 scores.

3.2.3 Supervised Fine-Tuning of GPT

We performed fine-tuning with the OpenAI API using the full labeled E-DAIC training set. The API performs parameter-efficient fine-tuning of pre-trained GPT weights, using entropy loss on the predicted tokens (in our case PHQ-8/PCL-5 scores). Our hypothesis was that, by increasing the number of examples seen and explicitly updating model weights, this approach could learn more nuanced patterns between VA and mental health assessment.

3.3 Datasets

3.3.1 GoEmotions

GoEmotions (Demszky et al., 2020) contains 58k curated Reddit comments in popular English, annotated by human raters with 27 fine-grained emotion labels. We used GoEmotions to fine-tune RoBERTa and enhance downstream fine-tuning for VA prediction.

3.3.2 EmoBank

EmoBank jackksoncsie (2022) is a corpus of 10k English sentences manually annotated with continuous Valence-Arousal-Dominance (VAD) scores. We used EmoBank to train and evaluate our RoBERTa-based VA regressors.

3.3.3 E-DAIC WOZ

The Extended Distress Analysis Interview Corpus - Wizard of Oz (E-DAIC) Gratch et al. (2014) Ringeval et al. (2019) is a multi-modal dataset of clinical interviews in which participants responded to scripted prompts. Participants are labelled with scores on self-administered PHQ-8 and PCL-5 assessments for depression and PTSD, respectively. E-DAIC is a gold-standard dataset for establishing benchmarks in automated mental health assessment. We used the text modality of interviews to evaluate text-to-depression and text-to-PTSD predictions using GPT.

4 Experiments

4.1 Experiments on RoBERTa-based VA prediction

Mean absolute error (MAE) was used to evaluate predictions of VA. Lower MAE indicates more accurate predictions in the two-dimensional VA space.

Experiments on training hyperparameters: We initially trained the model with a frozen RoBERTa encoder, using a learning rate of 3×10^{-5}, batch size of 16, and 20 epochs. Since both training and validation MAE plateaued early, we introduced dynamic unfreezing of the encoder after a number of epochs if validation MAE failed to improve by more than a chosen threshold. We experimented with varying patience to unfreezing and thresholds for MAE improvement. This strategy was able to reduce MAE in subsequent epochs but eventually led to overfitting, as evidenced by decreasing training MAE and increasing validation MAE, likely because unfreezing exposed a large number of parameters that began fitting to noise in the EmoBank data. To counter this, we experimented with lower learning rates and gradient clipping post-unfreezing, which produced the best training regime across model variants.

Experiments on model architectures: We considered a range of architectures. Specifically, we experimented with (1) the number of fully connected (FF) layers appended to the RoBERTa encoder, (2) the dimensionality of the last hidden layer (32, 64, 128, 256), and (3) whether to use a single model or separate models to predict valence and arousal. For both the emotion-aware and base RoBERTa, the best performance was achieved with three FF layers and a hidden size of 128. Interestingly, separate models for valence and arousal resulted in higher overall MAE, suggesting that joint prediction benefits from shared representations.

4.2 Prompt-Length Sensitivity in PHQ-8 Prediction

Large prompts can obscure clinically salient cues (Levy et al., 2024). We experimented with four strategies to compress transcripts Across all settings (see B).

Exp. 1: Transcript Summarization. We tasked GPT to summarize transcripts, specifically keeping clinically salient lines (with sentence-by-sentence VA) and condensed the rest (with average VA) (see C, and D). Both base and AutoCOT models showed higher MAE overall, suggesting that summary noise distorted both rationales and final predictions.

Exp. 2: Length-Based Pruning. We pruned transcripts by dropping utterances shorter than the transcript's mean line length. The result (see E) was that MAE increased, indicating that even brief replies (e.g. "I'm tired", or "not really") may encode important depressive cues.

Exp. 3: VA-Based Pruning. We removed lines outside of the 25–75th-percentile VA range for patients with depression or PTSD (see, F). Performance again declined, which suggests that *dynamics* of VA rather than extremes might be interpreted by GPT-based models.

Exp. 4: Number of AutoCoT Demos. We varied the number of in-context demos in AutoCOT (2, 4, and 6 demos). Interestingly, increasing the number of demos did not enhance performance, except in the case of length-pruned transcripts (see, G) where demos were likely recovering partially lost context.

5 Results and Discussion

5.1 Valence-Arousal Prediction

Table 1: Performance of Valence-Arousal prediction models on EmoBank test set

Metric	GPT-o4-mini	RoBERTa w/ Emotional Context	RoBERTa Direct for VA
MAE	0.197	0.0792	0.0759
Valence Pearson	0.789	0.7893	0.8201
Arousal Pearson	0.429	0.5479	0.5748

We evaluate performance with test-set mean absolute error (MAE) and the Pearson correlation between predicted and gold valence–arousal (VA) scores. The base RoBERTa regressor attains the best mean-squared error (MSE $= 0.0759$ in the $[-1, 1]$ VA space) and is therefore carried forward in subsequent mental-health experiments. Its high valence correlation indicates reliable separation of positive and negative language, while arousal remains harder to predict. Even before encoder unfreezing, the base RoBERTa model slightly outperforms an emotion-aware GoEmotions counterpart (MAE 0.0895 vs. 0.0946), suggesting that explicit emotion context neither speeds convergence nor sharpens VA discrimination. A zero-shot GPT-o4-mini baseline scores lowest on all metrics, underscoring the need for fine-tuning with VA-annotated examples.

5.2 PHQ-8

Table 2: PHQ-8 prediction on baseline transcripts

Metric	Base GPT	AutoCoT	Supervised Fine-Tuned GPT
Baseline — Without VA			
MAE	4.001	4.427	3.661
RMSE	5.513	6.257	4.520
With VA			
MAE	3.768	3.938	3.214
RMSE	5.174	5.437	4.520

Table 3: Best MAE Achieved by Each Model Across All Experiments

Model	**Condition**	**Best MAE**
Base GPT	With VA, baseline transcripts	4.036
AutoCoT	Without VA, baseline transcripts $(n = 4)$	3.916
Supervised Fine-Tuned GPT	With VA, baseline transcripts	3.214

Continuous valence–arousal (VA) features consistently lower error (Table 2): Base GPT MAE ↓ 5.8% and RMSE ↓ 6.1%, AutoCoT MAE ↓ 11.0% and RMSE ↓ 13.1%, and supervised fine-tuned GPT MAE ↓ 12.2% with no RMSE penalty. Even in strong few-shot settings, the fine-tuned + VA model remains best (MAE 3.214; Table 3).

We tested GPT summarization as a baseline compression method (see Appendix D). Although adding VA improved summaries by 9.3%, these VA-enhanced summaries still underperformed VA-augmented full transcripts, indicating that compression loses diagnostic information. Even our best summarization approach fell short of the supervised fine-tuned model, confirming that complete context and task-specific training are essential.

Our core finding is that continuous valence–arousal (VA) features provide a lightweight yet consistently informative signal for PHQ-8 prediction from text, modestly improving PHQ-8 estimation across GPT-based models. These improvements demonstrate that even a low-dimensional affective signal can supplement linguistic content when predicting depressive-symptom severity in large foundational models.

Another important finding is that most pruning strategies—based on length, summarization, or VA thresholds—uniformly degrade performance. This suggests three insights:

1. **Preserve conversational flow.** Short utterances and neutral lines still carry diagnostic cues (pauses, hesitations, shifts in affect) that VA amplifies, but that are lost when context is truncated.
2. **Quality over quantity of demonstrations.** In the AutoCoT few-shot pipeline, reducing to four high-quality demos outperformed larger sets, underscoring the fragility of rationale-based prompting in clinical tasks.
3. **Modeling trade-offs.** While supervised fine-tuning yields the best overall performance, VA-augmented zero- and few-shot methods remain competitive, offering rapid deployment when labeled data are scarce.

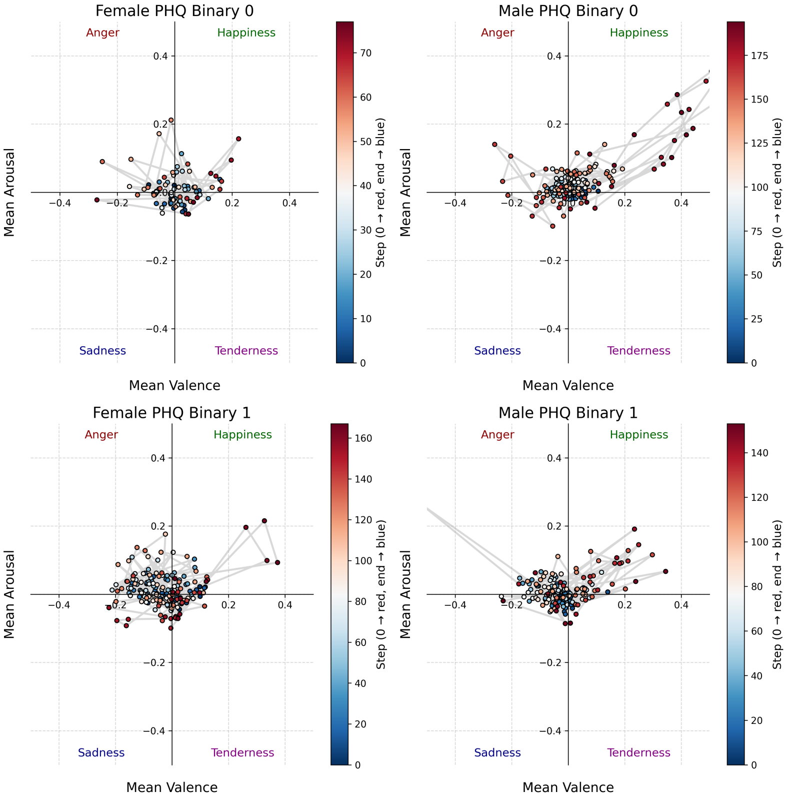

Figure 4: Valence-arousal trajectories by gender and depression status. Color progression from early (red) to late (blue) conversation evolution. Depressed groups show more negative valence excursions.

Even short, neutral exchanges carry subtle linguistic and affective cues—hesitations, tone shifts, self-reflective remarks—that the network uses to gauge depressive severity. Clinically, removing these fragments erases important indicators of mood fluctuation and engagement, limiting the model's ability to form a holistic understanding of the patient's emotional state. Future work should therefore preserve the full transcript, including interviewer prompts, to capture complete context and maintain conversational flow.

AutoCoT investigations revealed no consistent accuracy gains when varying demonstrations from two to eight, except for a slight improvement in length-pruned conditions ($4.952 \rightarrow 4.454$ MAE). The best AutoCoT performance was achieved with just four carefully curated demos, underscoring that clinical coherence and example quality matter far more than sheer quantity.

Beyond diagnostics, VA could enable therapeutic chatbots to generate emotionally calibrated responses by defining target VA coordinates for different clinical scenarios. For instance, when users present with low valence and high arousal (anxiety), the chatbot could target responses at $valence = 0.2,\ arousal = -0.3$ to provide gentle reassurance while lowering activation. Conversely, for low-valence, low-arousal states (depression), a target of $valence = 0.1,\ arousal = 0.2$ could offer mild encouragement without overwhelming the user. The system would constrain token sampling during generation to maintain trajectories toward these therapeutic targets while continuously monitoring user VA to adjust dynamically, creating an adaptive feedback loop in which both the current emotional state and desired therapeutic direction guide response generation.

5.3 Exploration of VA patterns with PHQ-8

To explore how valence–arousal (VA) patterns evolve during clinical interviews, we visualized average VA trajectories across conversation steps for different demographic and diagnostic groups. Figure 4 presents these trajectories in two-dimensional space with color-coded progression, whereas Figure 6 shows separate time-series plots of valence and arousal.

Figure 4 reveals distinct group-level patterns. Non-depressed participants remain largely in neutral territory: females drift toward higher valence, and males shift toward higher arousal and positive valence by the interview's end.

Depressed participants exhibit markedly different dynamics. Females enter negative-valence territory early and remain in low-arousal, negative regions, whereas males display fluctuating paths with negative-valence excursions at higher arousal. Both depressed groups cluster in the third quadrant (negative valence, negative arousal), corresponding to sadness in VA space. More detailed discussion is in Appendix H.

These trajectories demonstrate that VA dimensions capture meaningful emotional dynamics differentiating depressed from non-depressed individuals. Depression is characterized by lower average valence and distinctive temporal patterns during clinical interactions, supporting VA's utility as a continuous monitoring metric for mental-health applications.

5.4 PTSD

Table 4: PCL-5 prediction on baseline transcripts

Metric	Base GPT	AutoCoT	Supervised Fine-Tuned GPT
Baseline — Without VA			
MAE	14.870	12.532	8.963
RMSE	18.068	16.712	13.077
Accuracy (%)	62.5	66.3	72.7
With VA			
MAE	15.833	14.120	9.648
RMSE	18.893	17.812	14.566
Accuracy (%)	73.2	68.2	78.0
Random noise as VA			
MAE	15.759		
RMSE	20.279		
Accuracy (%)	67.6		

The addition of VA to transcripts improved the accuracy of binary PTSD predictions across all GPT-based models in our experiments (Table 4). The most significant improvement was seen with Base GPT where adding VA improved accuracy by 10.7%. We considered the possibility that the improvement might be due to VA values injecting random noise into the prompt, which has been shown in some cases to improve LLM performance by reducing bias in the attention mechanism (Zhao et al., 2021). However, when we presented random noise in the range [-1, 1] as VA, we saw just a 5.1% improvement in accuracy, and therefore semantic understanding of VA was likely improving the prediction of self-reported PTSD. This is supported by psychiatric studies on PTSD showing an association between VA (self-reported or measured by fMRI) and PTSD symptom severity in patients (Shin et al., 2005)(Lanius et al., 2016).

Figure 7 shows that non-PTSD participants of both genders maintain near-neutral valence and relatively flat arousal, whereas PTSD-positive individuals exhibit more frequent negative valence dips and larger arousal swings, most pronounced in males. Detailed discussion is in the appendix.

That said, we were surprised that the addition of VA did not improve MAE or RMSE across the GPT-based models and was usually associated with higher error than transcripts alone. We observed that for transcripts with VA, the models were responding with a small set of PCL-5 scores, while for transcripts alone, the models would output across a larger range of scores in the PCL-5 scale. These findings, together with the accuracy metrics on the binary classification task, suggest that VA might be more useful for the prediction of few, discrete classes. Future studies could explore if calibrating mental health scales, for example discretizing scales using thresholds, could improve the utility of VA.

We hypothesized that few-shot prompting with AutoCoT could help the GPT model develop insight into how VA relates to PTSD, but surprisingly the addition of VA features provided just a 1.9% accuracy improvement and a worsening of MAE and RMSE on PCL-5. This is contrary to our results on AutoCoT for depression where VA reduced MAE and RMSE of PHQ-8 scores. We experimented with AutoCoT few-shot prompting using summarized, length-pruned, and VA-pruned transcripts (Appendix I) as highlighted in section 4, but none were able to improve the utility of VA. Therefore, the benefit of few-shot prompting for learning VA features might depend on the disease context.

Supervised fine-tuned (SFT) GPT on the DAIC-WOZ train set achieved the highest performance in both PCL-5 score and accuracy, followed by AutoCoT and Base GPT. This makes sense given the SFT model had seen the highest number of examples and had its weights explicitly updated. Although our goal was to evaluate performance on prompts with VA versus text alone, we were impressed with the absolute performance of the SFT model with VA features, which exceeds the best reported accuracy (75.8%) for GPT-o4-mini (Ali et al., 2024) and achieves similar accuracy to domain-specific models based on random forest regression on sentiment features (Sawalha et al., 2022). Future studies could evaluate if supervised fine-tuning can preserve the foundational ability of GPT across mental health conditions – starting with how well the PTSD fine-tuned GPT may transfer to prediction of depression scores and vice versa.

6 Conclusion

We introduced a unified, two-dimensional valence–arousal framework that generalizes across depression and PTSD prediction tasks without disorder-specific retraining. Our best RoBERTa regressor achieved 0.0759 MAE on continuous VA estimation, and augmenting clinical transcripts with these affective signals yielded up to 12.2% MAE reduction for PHQ-8 and 10.7% accuracy gain for PTSD classification. Temporal VA trajectories further revealed distinct emotional dynamics in depressed and PTSD-positive individuals, underscoring the value of continuous affect monitoring. By decoupling affective representation from specific diagnoses, our approach paves the way for scalable, real-time mental health screening and personalized intervention in diverse clinical settings.

7 Contributions

- **Miko** formulated the psychiatric framing, identified valence–arousal (VA) as a foundational signal, and ran the PHQ-8 prediction experiments with VA input.
- **Kevina** led prompt-engineering studies, analysing VA sensitivity to transcript length and content.

- **Christo** developed the VA-regression models, contrasting GoEmotions and EmoBank pretraining regimes.
- **Yalcin** processed experimental outputs and designed the 2-D VA-trajectory visualisations of depressive progression.
- **Ion** implemented the PTSD benchmark, demonstrating VA's utility across two clinical conditions and broadening the study's scope.

All authors discussed the results, contributed to the manuscript, and approved the final submission.

References

Abdelrahman A. Ali, Aya E. Fouda, Radwa J. Hanafy, and Mohammed E. Fouda. Leveraging audio and text modalities in mental health: A study of LLMs performance. *arXiv preprint arXiv:2412.10417*, December 2024. URL `https://arxiv.org/abs/2412.10417v1`. version v1, submitted Dec. 9 2024.

Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1): 49–59, 1994. ISSN 0005-7916. doi: https://doi.org/10.1016/0005-7916(94)90063-9. URL `https://www.sciencedirect.com/science/article/pii/0005791694900639`.

Sergio Burdisso, Ernesto A. Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, A. Pastor López-Monroy, and Petr Motlicek. Daic-woz: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. *arXiv preprint arXiv:2404.14463*, v1, April 2024. URL `https://arxiv.org/abs/2404.14463v1`. License: arXiv.org perpetual non-exclusive license.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.aclâĂŚmain.372. URL `https://aclanthology.org/2020.acl-main.372/`.

Roger Garriga, Javier Mas, Semhar Abraha, Jon Nolan, Oliver Harrison, George Tadros, and Aleksandar Matic. Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*, 28(6):1240–1248, 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01811-5. URL `https://doi.org/10.1038/s41591-022-01811-5`.

Global Burden of Disease 2017 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1789–1858, 2018. ISSN 0140-6736. doi: 10.1016/S0140-6736(18)32279-7. URL `https://www.sciencedirect.com/science/article/pii/S0140673618322797`.

Jonathan Gratch, Ron Artstein, Gale M. Lucas, Georgia Stratou, Stefan Scherer, Arno Nazarian, Robin Wood, Jill Boberg, David DeVault, Stacy Marsella, and David R. Traum. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128, Reykjavik, Iceland, May 2014.

Yanrong Guo, Jilong Liu, Lei Wang, Wei Qin, Shijie Hao, and Richang Hong. A prompt-based topic-modeling method for depression detection on low-resource data. *IEEE Transactions on Computational Social Systems*, 11(1):1430–1439, 2024. doi: 10.1109/TCSS.2023.3260080.

jackksoncsie. Emobank: A corpus of 10k english sentences annotated with emotion (vad). `https://www.kaggle.com/datasets/jackksoncsie/emobank`, 2022. Kaggle dataset. Licensed under CC-BY-SA 4.0. Accessed: 2025-06-08.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*, 2021. `https://arxiv.org/abs/2110.15621`.

Jiehui Jia, Huan Zhang, and Jinhua Liang. Bridging discrete and continuous: A multimodal strategy for complex emotion detection. *arXiv preprint arXiv:2409.07901*, v1, September 2024. URL https://arxiv.org/abs/2409.07901v1. License: CC BY 4.0.

Arun Kashyap. Mental health chatbot using roberta and gemini. https://github.com/kashyaparun25/Mental-Health-Chatbot-using-RoBERTa-and-Gemini, 2024. GitHub repository. Accessed: 2025-06-08.

Ruth A. Lanius, Paul A. Frewen, Annie Vermette, Sarah Myers, Bernhard Brand, Margaret C. McKinnon, Regan W. J. Neufeld, Britt Schuhmann, and Keith T. Brady. Emotional processing in posttraumatic stress disorder: Heightened negative emotionality and related emotional numbing. *BMC Psychiatry*, 16:24, 2016. doi: 10.1186/s12888-016-0770-3. PMC4843505.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models, 2024. URL https://arxiv.org/abs/2402.14848.

Michael Mitsios, Georgios Vamvoukakis, Georgia Maniati, Nikolaos Ellinas, Georgios Dimitriou, Konstantinos Markopoulos, Panos Kakoulidis, Alexandra Vioni, Myrsini Christidou, Junkwang Oh, Gunu Jho, Inchul Hwang aand Georgios Vardaxoglou, Aimilios Chalamandaris, Pirros Tsiakoulis, and Spyros Raptis. Improved text emotion prediction using combined valence and arousal ordinal classification, 2024. URL https://arxiv.org/abs/2404.01805.

Alireza Pourkeyvan, Ramin Safa, and Ali Sorourkhah. Harnessing the power of hugging face transformers for predicting mental health disorders in social networks. *IEEE Access*, 12: 28025–28035, 2024. ISSN 2169-3536. doi: 10.1109/access.2024.3366653. URL http://dx.doi.org/10.1109/ACCESS.2024.3366653.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Denis Lalanne, Maja Pantic, Mihalis A. Nicolaou, et al. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12. ACM, 2019. doi: 10.1145/3347320.3357688.

Misha Sadeghi, Robert Richer, Bernhard Egger, Lena Schindler-Gmelch, Lydia Helene Rupp, Farnaz Rahimi, Matthias Berking, and Bjoern M. Eskofier. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, 3:66, 2024. doi: 10.1038/s44184-024-00112-8. URL https://www.nature.com/articles/s44184-024-00112-8. Published 23 December 2024, License: CC BY 4.0.

Jeff Sawalha, Muhammad Yousefnezhad, Zehra Shah, Matthew R. G. Brown, Andrew J. Greenshaw, and Russell Greiner. Detecting presence of ptsd using sentiment analysis from text data. *Frontiers in Psychiatry*, 12:811392, February 2022. doi: 10.3389/fpsyt.2021.811392. URL https://www.frontiersin.org/articles/10.3389/fpsyt.2021.811392/full.

Lisa M. Shin, Christopher I. Wright, Paul A. Cannistraro, Michelle M. Wedig, Katherine McMullin, Brian Martis, Michael L. Macklin, Natasha B. Lasko, Sarah R. Cavanagh, Terri S. Krangel, Scott P. Orr, Roger K. Pitman, Paul J. Whalen, and Scott L. Rauch. A functional magnetic resonance imaging study of amygdala and medial prefrontal cortex responses to overtly presented fearful faces in posttraumatic stress disorder. *Archives of General Psychiatry*, 62(3):273–281, March 2005. doi: 10.1001/archpsyc.62.3.273. PMID: 15753240.

Ala N. Tak and Jonathan Gratch. Gpt-4 emulates average-human emotional cognition from a third-person perspective. *arXiv preprint arXiv:2408.13718*, v1, August 2024. URL https://arxiv.org/abs/2408.13718v1. License: CC BY 4.0.

Yuqing Wang, Yun Zhao, Sara Alessandra Keller, Anne de Hond, Marieke M. van Buchem, Malvika Pillai, and Tina Hernandez-Boussard. Unveiling and mitigating bias in mental health analysis with large language models. *arXiv preprint arXiv:2406.12033*, v2, June 2024. URL https://arxiv.org/abs/2406.12033v2. License: CC BY-NC-SA 4.0.

Zhiting Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*, 2021. URL https://arxiv.org/abs/2102.09690.

Michael Zimmerman. The value and limitations of self-administered questionnaires in clinical practice and epidemiological studies. *World Psychiatry*, 23(2):210–212, 2024. doi: 10.1002/wps.21191. URL `https://doi.org/10.1002/wps.21191`.

A AutoCOT Rationale Generation Prompt

```
You are a clinical psychologist.  You read transcripts of a patient
from a diagnostic interview and estimate the PHQ-8 score (0-24) by
reasoning through what the participant said.  You will be given what
the patient said line by line.  Return only your reasoning - do not
include the final score.
```

In the VA case, the prompt is edited as follows:

```
You will be given what the patient said line by line, along with the
valence and arousal score (-1 to 1) for that line.
```

Specifically, we use the following prompt template for each transcript:

```
N=8 demo prompt.
Q: You are a clinical psychologist.  You read transcripts of a
patient from a diagnostic interview and estimate the PHQ-8 score
(0-24) by reasoning through what the participant said.  You will be
given what the patient said line by line
Evaluate the patient's total PHQ8 score
{formatted test transcript}
A: Let's think step by step.  After analyzing the transcript, you
MUST output the score in this exact format:
PHQ_8Total:  [score]
Again, make sure to output the score in the above format.
```

Again, In the VA case, the prompt is edited as follows:

```
You will be given what the patient said line by line, along with the
valence and arousal score (-1 to 1) for that line.
```

B Summary of All AutoCOT Experiments

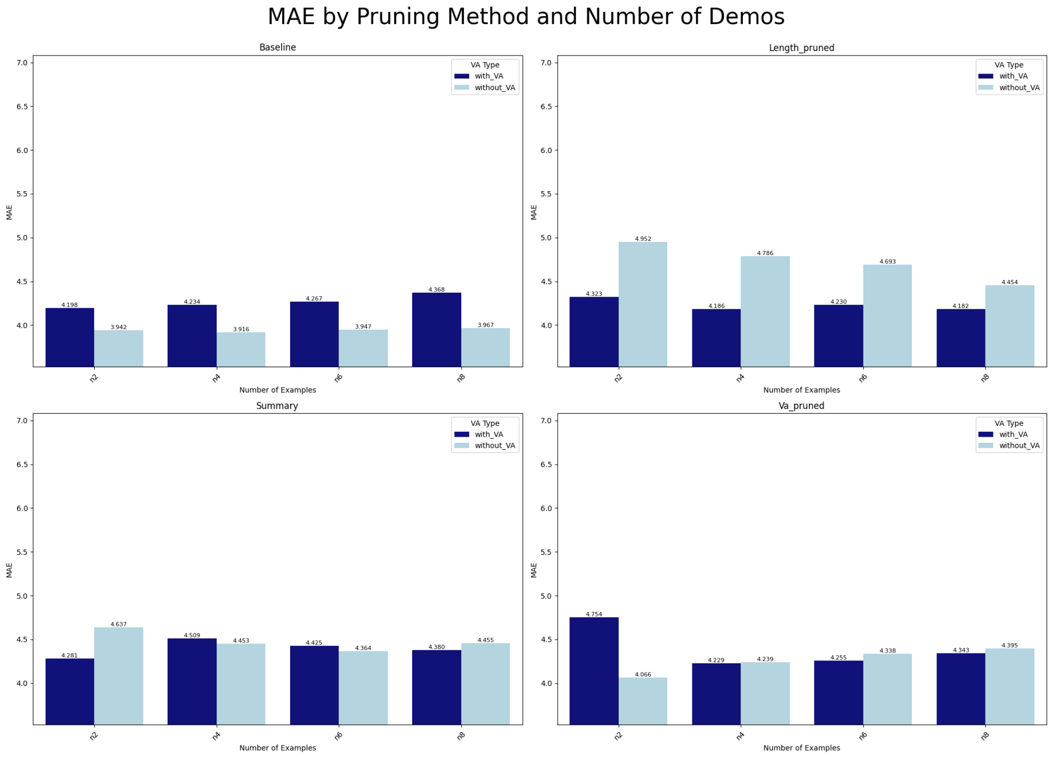

Figure 5: Aggregation of PHQ-8 MAE AutoCOT variations across 1) with or without VA 2) Pruning mechanism — baseline, summary, length, va, and 4) Number of demos — n = 2, 4, 6, 8.

C Experiment 1: Summary — Example summarized transcript

Example Utterance with VA (Verbatim):
‘‘I don’t know I’m just in a happy mood today’’
Valence: 0.3847 *Arousal:* 0.1264

Summary Line with VA:
‘‘The transcript contains many brief acknowledgments and casual comments (e.g., ‘okay,‘ ‘yes,‘ comments about Mexico, weather, crime, music, Burger). The client reports writing poetry and aspirations to become a famous poetry writer. She dropped out in 11th grade. About a week ago she had an argument with her husband over cigarettes; he wants to quit smoking and she doesn’t. She has a history of drug addiction, received help from her husband, feels better now, and previously overdosed on pills. She is not participating in therapy at this time and is uncomfortable discussing some topics. She experiences hypersomnia (‘sleep too much‘) and uses music and erotic poetry as coping strategies. Overall, she describes feeling happier, more outgoing, occasionally cranky, and in a generally positive mood today.’’
Valence: 0.0414 *Arousal:* -0.0075

D Experiment 1: Summary — Results

Table 5: PHQ-8 prediction using summary transcripts: Base GPT vs. AutoCoT

Metric	Base GPT	AutoCoT
Without VA — Baseline		
MAE	4.667	6.000
RMSE	5.994	7.330
With VA		
MAE	**4.036**	4.500
RMSE	5.379	6.619

E Experiment 2: Prune by Length — Results

Table 6: PHQ-8 prediction using length-pruned transcripts: Base GPT vs. AutoCoT

Metric	Base GPT	AutoCoT
Without VA — Baseline		
MAE	4.107	4.454
RMSE	**5.369**	**6.052**
With VA		
MAE	4.125	4.182
RMSE	5.433	5.788

F Experiment 3: Prune on VA — Results

Table 7: PHQ-8 prediction using VA-pruned transcripts: Base GPT vs. AutoCoT

Metric	Base GPT	AutoCoT
Without VA — Baseline		
MAE	5.679	4.395
RMSE	7.711	6.144
With VA		
MAE	5.696	4.343
RMSE	7.784	5.846

G Experiment 4: Vary Number of Demos — Results

Table 8: PHQ-8 prediction across varying number of demos in AutoCOT

# Demo Texts	AutoCoT (without VA)		AutoCoT (with VA)	
	MAE	RMSE	MAE	RMSE
2	3.942	5.645	4.198	5.829
4	3.916	5.637	4.234	5.872
6	3.947	5.683	4.267	5.999
8	3.967	5.663	4.368	6.067

H Visualization of Dynamics of VA over Time Steps

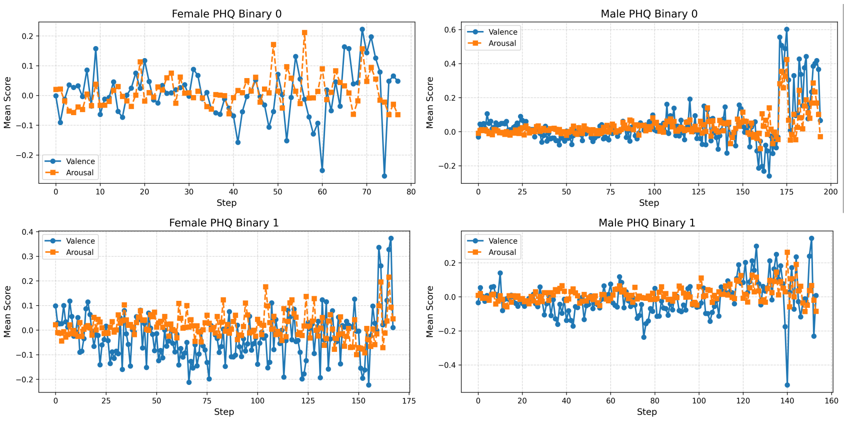

Figure 6: Temporal evolution of mean valence (blue) and arousal (orange) throughout clinical conversations by demographic-diagnostic groups. Depressed participants show more volatile patterns with frequent negative valence dips.

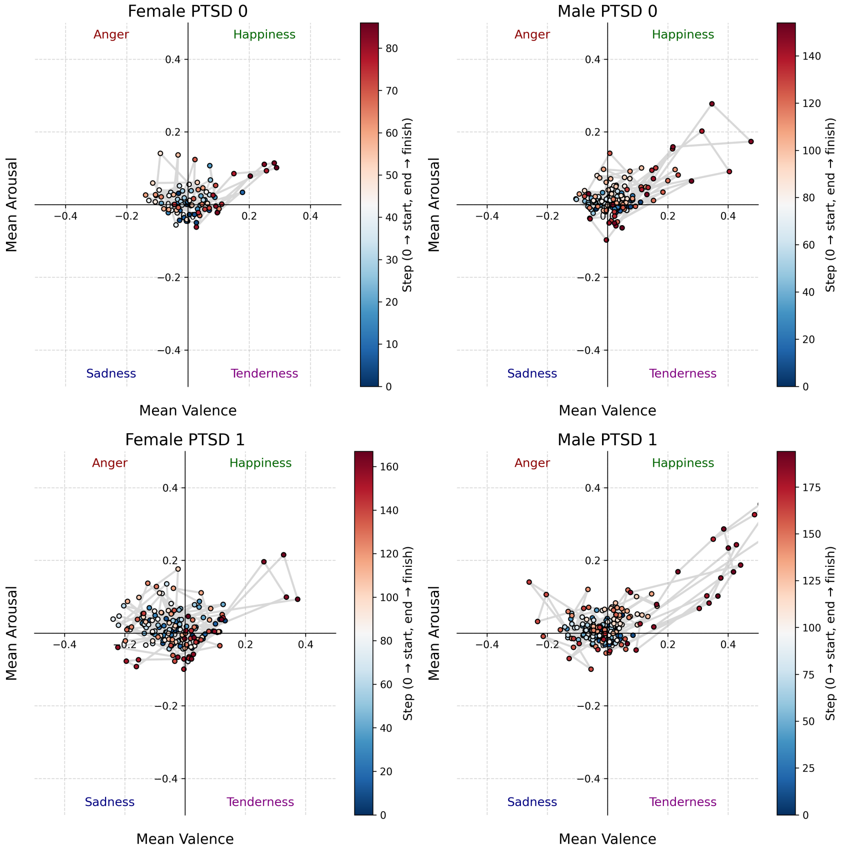

Figure 7: Evolution of mean valence (blue) and arousal (orange) over interview steps for participants with and without PTSD, separated by gender. PTSD-positive individuals exhibit more pronounced negative valence swings and greater arousal variability compared to their non-PTSD counterparts.

In PHQ8 dynamics, Figure 4 presents VA trajectories as paths through the two-dimensional emotional space, with colors indicating progression from early (red) to late (blue) conversation steps. Several notable patterns emerge when comparing groups stratified by gender and depression status.

Non-depressed participants (PHQ Binary 0) show relatively contained trajectories that generally remain within neutral emotional territory. The female non-depressed group exhibits a trajectory that starts near neutral and moves slightly toward positive valence regions as the conversation progresses. The male non-depressed group demonstrates a more dramatic shift, with the trajectory moving from neutral territory toward higher arousal and positive valence regions by the end of the interview.

Depressed participants (PHQ Binary 1) display clearly different trajectory patterns. The female depressed group shows a path that moves into negative valence territory early in the conversation (red points) and tends to remain in lower-arousal, negative-valence regions throughout the interaction. The male depressed group exhibits a more fluctuating trajectory with several excursions into negative valence regions, although with somewhat higher arousal levels than their female counterparts. A common pattern seen in both male and female depressed groups is dense clustering around the third quadrant (negative valence, negative arousal), which corresponds to the sadness region in VA categorization.

Figure 6 provides a clearer view of how valence and arousal change over the course of conversations. The step-by-step analysis reveals distinct temporal patterns.

For non-depressed females, valence fluctuates around neutral with occasional positive peaks, while arousal remains relatively stable with moderate variability. Non-depressed males show a striking pattern where both valence and arousal remain relatively stable throughout most of the conversation but show dramatic increases toward the final steps, potentially reflecting positive engagement or relief as the interview concludes.

Depressed females demonstrate consistently lower and more volatile valence patterns, with frequent dips into negative territory throughout the conversation. Their arousal levels remain relatively stable but tend toward the lower end of the range. Depressed males show more fluctuating valence patterns with sharp negative changes, while their arousal levels demonstrate greater variability compared to depressed females.

Figure 7 illustrates how emotional tone and activation shift throughout clinical interviews in relation to PTSD status and gender.

Non-PTSD females maintain valence close to neutral, with only minor positive and negative fluctuations; their arousal trace remains relatively flat, suggesting steady engagement without marked emotional peaks. Non-PTSD males similarly display stable valence, though with a subtle upward drift in the later steps, and a gently rising arousal curve toward the end of the conversation, perhaps reflecting growing comfort or relief.

By contrast, PTSD-positive females show recurrent drops into negative valence, interspersed with brief recoveries; their arousal profile also oscillates more sharply, indicating intermittent spikes of physiological activation. PTSD-positive males demonstrate the most volatile patterns: valence swings widely between neutral and negative, and arousal exhibits high amplitude peaks and troughs across the interview. These dynamics suggest that PTSD is associated with less regulated affective responses and heightened emotional reactivity during clinical dialogue.