Emotion Probes for Depression and PTSD in Clinical Interviews

Anonymous ACL submission

Abstract

Depression and PTSD detection from clinical interviews is usually framed as a generic text classification problem, yet what makes these interviews clinically informative is affective signal that often carries patterns participants do not openly express. We propose an emotion probe pipeline that reads this signal directly from the internal representations of a frozen LLM. The method calibrates 28 emotion directions from paired emotional and neutral message prompts, projects per-turn hidden states from Mistral 24B onto these directions, and fits a small linear classifier to predict depression and PTSD severity. Across DAIC-WOZ depression, E-DAIC depression, and E-DAIC PTSD, our method achieves the highest AUROC among text-only approaches, outperforming six baselines spanning sentiment analysis, fine-tuned BERT, pretrained emotion classifiers, LLM prompting, and clinical scoring. Adapting to a new condition requires only a new topic list rather than additional clinical labels. Our method also runs 1.4× faster than LLM prompting and 4.4× faster than clinical scoring on the same backbone. Feature selection recovers compact, clinically interpretable emotion sets that align with established symptom patterns for each condition.

1 Introduction

Depression and PTSD are common psychiatric conditions that remain difficult to detect reliably in clinical settings. Standard screening instruments such as the PHQ-9 for depression (Kroenke et al., 2009) and the PCL-C for PTSD (Blanchard et al., 1996) produce structured numeric assessments but rely on self-report, which may not fully capture a patient's internal state. Semi-structured clinical interviews complement self-report by examining how a person speaks about their experience. The DAIC-WOZ dataset brought this approach to scale by pairing Wizard-of-Oz virtual interviews with PHQ-8 labels, and its extension E-DAIC adds PCL-C PTSD labels alongside a fully automated interviewer (Gratch et al., 2014; Ringeval et al., 2019). The shift toward automated interviewers tracks a broader move in conversational AI for mental health, where large language models are increasingly used to conduct, support, or analyze clinical conversations (Na et al., 2025).

Existing methods for depression and PTSD detection from clinical interviews fall into three families, each typically tied to a single clinical outcome. Supervised models fine-tune text or multimodal encoders end-to-end on the target label (Burdisso et al., 2023; Al Hanai et al., 2018). Direct LLM prompting asks the model to produce a diagnosis in zero-shot or few-shot form (Yang et al., 2024), drawing on general clinical knowledge but providing little transparency into what drove the decision. Clinical scoring methods such as AIDA (Lee et al., 2026) prompt an LLM for a fixed set of clinically grounded item scores, recovering transparency through prompt engineering. All three families treat the clinical interview as a generic text classification problem. Yet what makes interviews diagnostically useful is affective: depression is marked by sustained negative affect and reduced positive affect (Watson et al., 1988), and PTSD by intrusive distress, heightened threat response, and difficulty feeling positive emotions. Prior work has tried to capture this by feeding outputs of a pretrained emotion classifier into a downstream model (Wu et al., 2022; Xezonaki et al., 2020), but surface-level emotion classifiers miss the clinically informative patterns participants do not openly express, including suppression, withdrawal, and rumination.

Pretrained emotion classifiers operate only on what a participant literally says, so they cannot detect a participant who responds "I'm doing okay, I guess" while internally experiencing the flat affect characteristic of depression. We argue that the right representation for this problem lives in-

side the language model rather than in its outputs. Recent interpretability research shows that LLMs encode emotion as nearly linear directions in activation space, and that these directions respond to implicit emotional content even when no explicit emotional language is present (Zou et al., 2023; Sofroniew et al., 2026). We calibrate 28 emotion directions from paired emotional and neutral message prompts, then project per-turn hidden states from Mistral Small 3.2 24B Instruct at each participant response boundary onto these directions to obtain a 28-dimensional emotion vector per turn. We average across turns to obtain a per-participant feature and fit a small linear classifier that maps this feature to a depression or PTSD prediction. We evaluate the approach on DAIC-WOZ and E-DAIC across depression and PTSD and compare against six baselines.

Our contributions are:

- We propose an emotion probe pipeline for depression and PTSD prediction that requires no clinical labels at calibration time, where adapting to a new condition only needs a new topic list rather than relabeling, prompt tuning, or retraining.
- Across DAIC-WOZ depression, E-DAIC depression, and E-DAIC PTSD, our method achieves the best AUROC among text-only methods (0.846, 0.836, 0.846), beating six baselines on the same Mistral 24B backbone.
- We compare inference cost with baselines and show that our method runs 1.4× faster than direct LLM prompting and 4.4× faster than clinical scoring method.
- Feature analysis recovers clinically interpretable emotion sets for each condition.

2 Related Work

Text-based depression and PTSD detection. DAIC-WOZ (Gratch et al., 2014) and its extension E-DAIC are the standard benchmarks for automated depression and PTSD detection from clinical interviews, providing semi-structured Ellie interviews with PHQ-8 labels and, in E-DAIC, additional PCL-C PTSD labels. Because clinical interview data is expensive and rarely shareable, this well-documented corpus with official train/dev/test splits has become the default benchmark, evaluated by a large literature spanning topic modeling (Gong and Poellabauer, 2017), multi-instance learning (Zhang et al., 2025), in-context learning (Zhao et al., 2025a), and supervised fine-tuning. A common challenge is that most interview text is conversational filler unrelated to the clinical outcome, so treating the whole interview as one input lets irrelevant content weaken the prediction (Zhang et al., 2025), while Burdisso et al. (2023) report the opposite failure of classifiers focusing too narrowly on the mental-health question region.

LLM-based methods for clinical interview assessment have grown rapidly (Rosenman et al., 2024; Na et al., 2025). One study uses direct zero-shot or few-shot prompting to predict severity scores from interview transcripts (Ohse et al., 2024; Kebe et al., 2025). Another study treats the LLM as an intermediate feature extractor. Rosenman et al. (2024) ask the model to respond to PHQ-8 and PCL-C questions as if it were the participant, then train a regression model on these responses. Lee et al. (2026) prompt the model with 23 clinically grounded questions and feed the resulting factor scores into a linear regression model. Recent extensions add multi-agent reasoning and audiovisual fusion on top of LLM backbones (Zhao et al., 2025b). Unlike these methods, which extract clinical features through text generation, our method reads them directly from the LLM's internal activations.

Mechanistic interpretability of emotion representations. Recent work on LLM interpretability shows that LLMs encode high-level concepts as nearly linear directions in their activation space, recoverable by contrasting activations on paired prompts (Zou et al., 2023). Li et al. (2024) show that emotion concepts are encoded in specific neurons and causally support emotion inference, Tak et al. (2025) localize these representations to mid-layer hidden states, and Sofroniew et al. (2026) find 171 emotion directions in Claude Sonnet 4.5 that causally shape downstream behavior. These findings suggest that emotion is not only encoded in LLM activations but actively shapes downstream generation, making internal emotion representations a candidate signal for any task where affective content drives the prediction.

We apply this perspective to a clinical setting. Clinicians assess depression and PTSD severity largely from the emotional content of what patients say but existing methods either pool text features across the whole interview or prompt the LLM to

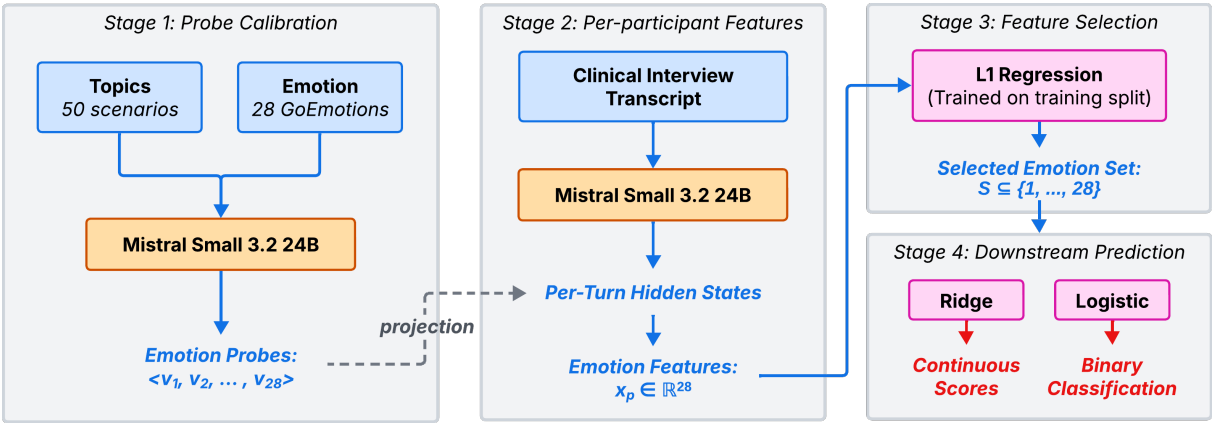

Figure 1: Pipeline overview.

generate scores which ignore the emotion signal the model already computes internally. We extract internal emotion directions and use them as features to predict depression and PTSD scores on the DAIC-WOZ and E-DAIC interview datasets. To our knowledge, this is the first work to use LLM internal emotion representations for depression and PTSD detection on a benchmark clinical interview dataset.

3 Method

Figure 1 summarizes the four-stage pipeline. The LLM stays frozen throughout; only the Lasso filter and a small linear head are trained on clinical labels. Full implementation details (model identifier, hyperparameters, hardware) are in Appendix A.

Backbone. We use Mistral Small 3.2 24B Instruct as the frozen backbone, with per-turn hidden states extracted at the last token of each participant turn.

Stage 1: Probe calibration. For each of the 28 GoEmotions categories (Demszky et al., 2020), we generate 50 first-person message pairs over a depression-relevant topic set (Appendix B): an emotional variant that implicitly conveys the emotion without naming it and a matched neutral variant on the same topic. Following representation-engineering practice (Zou et al., 2023), we extract per-turn hidden states for each message, average within each variant $(\bar{h}_j^{\text{emo}}, \bar{h}^{\text{neu}})$, and define the probe direction as the unit-normalized difference,

$$v_j = \frac{\bar{h}_j^{\text{emo}} - \bar{h}^{\text{neu}}}{\|\bar{h}_j^{\text{emo}} - \bar{h}^{\text{neu}}\|}, \tag{1}$$

with the source layer chosen per direction by a valence/arousal PCA quality score. For PTSD we recalibrate the same 28 directions on a parallel set of 50 PTSD-relevant topics organized around the four DSM-5 symptom clusters, holding the emotion vocabulary and every other component fixed. We additionally construct a combined calibration that merges all 100 depression and PTSD topics into a single set including both depressive and trauma-related content. Section 6.4 evaluates the three calibrations and shows that matching them to the clinical content of the corpus is essential for performance.

Stage 2: Per-participant features. Appendix C illustrates the projection at the finer token level on example passages from a depressed and a non-depressed participant, showing how each emotion direction activates differently across the two scenarios. For evaluation, we extract per-turn hidden states at every participant turn for each interview. Each per turn state h is projected onto every v_j via the inner product $\langle h, v_j \rangle$, outputting a 28-dimensional activation per turn. The participant-level feature $x_p \in \mathbb{R}^{28}$ is the mean across all turns.

Stage 3: Feature selection. The 28 emotion directions are not statistically independent, since GoEmotions categories like joy, amusement, and excitement, or sadness, grief, and disappointment, share substantial semantic and clinical structure, producing high pairwise correlations (Appendix D) and the multicollinearity. With only 107 to 163 training participants per dataset, we apply an L1-penalized (Lasso) regression on the training split to select a subset $S \subseteq \{1, \ldots, 28\}$ of directions with nonzero coefficient, since Lasso both prevents overfitting and produces a sparse subset of directions that absorbs the redundant content into a smaller representative set. The penalty λ is chosen by 5-

fold cross-validation. For evaluation in Table 1, we report results at 5, 10, and 18 selected features for the three settings.

Stage 4: Downstream prediction. The selected features $x_p[S]$ are standardized using training statistics and fed into two downstream models: Ridge regression for continuous severity (PHQ-8 or PCL-C) and Logistic regression for binary classification. For binary classification in depression, we use PHQ-8 ≥ 10, which identifies clinically significant depressive symptoms (Kroenke et al., 2009), and for PTSD we use PCL-C ≥ 44, the threshold commonly used to screen for PTSD diagnosis (Blanchard et al., 1996). Both models use 5-fold cross-validation on the training split.

4 Experiment Setup

4.1 Datasets

DAIC-WOZ provides 189 semi-structured clinical interviews between participants and Ellie, a Wizard-of-Oz virtual interviewer, with PHQ-8 self-report labels (range 0–24). Each interview lasts roughly 7 to 33 minutes. We follow the canonical AVEC 2017 splits (Ringeval et al., 2017): 107 train, 35 development, 47 test, with 14 of 47 test participants meeting the PHQ-8 ≥ 10 depression criterion. Transcripts are human gold-standard.

E-DAIC extends DAIC-WOZ to 275 participants (163 train, 56 dev, 56 test) with both PHQ depression and PCL-C PTSD labels, collected by a fully automated interviewer. Because the released transcripts contain only the participant side but our per-turn projection requires turn boundaries defined by Ellie's questions as well, we re-transcribe the raw audio using ElevenLabs with speaker diarization to recover both speakers.

4.2 Transcription Quality on E-DAIC

Because E-DAIC participant transcripts are produced by ElevenLabs Scribe (human transcripts are not available for the new participants), we validate transcription quality against the DAIC-WOZ gold transcripts on the 171 sessions present in both corpora. The corpus-level word error rate is 5.60% (bootstrap 95% CI $[5.07\%, 6.21\%]$), well within the regime (< 15 to 20%) where downstream sentiment and emotion tasks are typically preserved and our features further average over hundreds of utterances per session. This supports extending evaluation to E-DAIC's automated transcripts. More details are in Appendix E.

4.3 Baselines

We compare against five baselines. Four of them produce a feature vector per utterance, which we feed into the same Ridge and Logistic classifiers used in our pipeline so that the comparison is controlled to the choice of features. Implementation details for all baselines are in Appendix F.

Lexicon/pretrained features. VADER (Hutto and Gilbert, 2014) (rule-based sentiment polarity) and Sentence-BERT (Reimers and Gurevych, 2019) (general-purpose sentence embeddings) test whether emotion-specific calibration adds value over off-the-shelf features.

GoEmotions classifier. A RoBERTa classifier (Liu et al., 2019) fine-tuned on the same 28 GoEmotions categories (Demszky et al., 2020) as our probes. This is the closest pretrained emotion classifier sharing our taxonomy.

Fine-tuned BERT. We fine-tune BERT (Devlin et al., 2019) end-to-end on the DAIC-WOZ/E-DAIC training labels, and report it in two forms: as a direct classifier, and as a feature extractor using the [CLS] embedding. This serves as a small-data supervised baseline.

AIDA-style clinical scoring. A reproduction of Lee et al. (2026), which prompts an LLM for 23 clinically grounded scores per interview. We run it on both Mistral 24B and GPT-4o.

Direct LLM prompting. The LLM (Mistral 24B or GPT-4o) is asked for a single binary diagnosis, in zero-shot and few-shot variants, following recent work on LLM-based mental health classification (Yang et al., 2024). Because it outputs only a binary label, it is scored on classification metrics (F1, Precision, Recall, and AUROC) but not on continuous severity (MAE, R^2).

Evaluation protocol. We use the same evaluation pipeline for every method. Features are standardized to zero mean and unit variance using training-set statistics. For continuous severity scores, we fit a Ridge regression with the regularization strength chosen by 5-fold cross-validation, and report MAE and R^2. For binary screening, we fit a Logistic regression under the same cross-validation setup, weighting the depressed or PTSD-positive group more heavily so it contributes equally to the loss despite being the minority. Binary metrics (AUROC, Precision, Recall, F1) are computed for

Method	DAIC-WOZ Depression		E-DAIC Depression		E-DAIC PTSD	
	AUROC ↑	F1 ↑	AUROC ↑	F1 ↑	AUROC ↑	F1 ↑
Ours: GoEmotions-28 probes (Mistral 24B)	0.84	0.686	**0.836**	0.619	0.844	0.667
Ours: Selected probes (5 / 10 / 18 features)	**0.846**	**0.757**	0.813	0.634	**0.846**	0.694
Same backbone (Mistral 24B), different extraction						
AIDA-style 23 questions (Mistral 24B)	0.82	0.67	0.816	0.545	0.841	0.652
Mistral zero-shot prompting	0.71	0.60	0.649	0.514	0.762	0.708
Mistral few-shot (1/class)	0.69	0.57	0.800	**0.698**	0.719	0.678
Mistral few-shot (3/class)	0.67	0.55	0.550	0.261	0.519	0.303
Frontier model baselines (GPT-4o)						
AIDA-style 23 questions (GPT-4o)	0.78	0.73	0.804	0.596	0.838	**0.750**
GPT-4o zero-shot prompting	0.79	0.69	0.750	0.649	0.795	0.744
GPT-4o few-shot (1/class)	0.77	0.67	0.749	0.638	0.762	0.708
GPT-4o few-shot (3/class)	0.81	0.71	0.710	0.600	0.748	0.694
Pretrained and lexicon baselines						
Sentence-BERT (MiniLM, 384-dim)	0.73	0.50	0.673	0.514	0.603	0.512
GoEmotions classifier (RoBERTa)	0.67	0.44	0.799	0.667	0.810	0.711
VADER sentiment	0.74	0.57	0.765	0.560	0.740	0.596
Supervised fine-tuning						
Fine-tuned BERT (direct end-to-end head)	0.77	0.47	0.351	0.000	0.451	0.071
Fine-tuned BERT ([CLS] features)	0.63	0.35	0.465	0.148	0.474	0.389

Table 1: Depression and PTSD detection across three evaluation settings. Our two methods are the full 28 emotion probes and the selected subset of 5, 10, and 18 directions for each setting. DAIC-WOZ is evaluated on depression probes and E-DAIC is evaluated on combined probes. Best per column in **bold**.

the clinical group, defined as PHQ-8 ≥ 10 for depression and PCL-C ≥ 44 for PTSD, following AVEC convention (Ringeval et al., 2017). All models are evaluated on the official test split, with no hyperparameter tuning on test data.

5 Results

Table 1 reports classification metrics across three settings: DAIC-WOZ depression, E-DAIC depression, and E-DAIC PTSD. We use AUROC and F1 in the main table for direct comparison; continuous prediction metrics (MAE, R^2) are reported inline. The choice of calibration set follows from the analysis in Section 6.4: depression probes for DAIC-WOZ, and combined depression+PTSD probes for both E-DAIC tasks. Section 6.4 explains why this domain specific choice is necessary.

DAIC-WOZ depression. With depression probes, the full 28-probe variant reaches AUROC 0.84 (MAE 3.89, R^2 0.44), the strongest among text-only methods. Pruning to 5 directions (fear, approval, amusement, anger, curiosity) gives a further lift to AUROC 0.846 and F1 0.757 (MAE 3.73), the best AUROC and F1 in the table. The closest competitor is AIDA-style scoring on Mistral 24B (AUROC 0.82, F1 0.67, MAE 3.92, R^2 0.44). Among GPT-4o baselines, few-shot prompting with 3 examples per class reaches AUROC 0.81, which is the next-strongest AUROC overall.

E-DAIC depression. With combined depression+PTSD probes, the full 28-probe variant leads on AUROC (0.836), edging out AIDA-style scoring on the same Mistral backbone (0.816) and GPT-4o (0.804). The pruned variant trails the full version slightly on AUROC (0.813) but improves F1 from 0.619 to 0.634. Mistral one-shot prompting takes the top F1 (0.698), though this advantage does not extend to the 3-shot setting where context-budget truncation drops F1 to 0.261. The GoEmotions classifier is also competitive (AUROC 0.799, F1 0.667), which indicates that E-DAIC interviews carry surface affective cues.

E-DAIC PTSD. With the same combined probes, the pruned variant achieves the highest AUROC (0.846), with the full 28-probe variant right behind (0.844). AIDA-style scoring is the next strongest (0.841 Mistral, 0.838 GPT-4o). On F1, AIDA-style scoring with GPT-4o leads (0.750), but with the same Mistral 24B backbone our pruned variant (0.694) beats AIDA-Mistral (0.652). The GoEmotions classifier remains strong here as well (AUROC 0.810, F1 0.711), showing that E-DAIC interviews contain surface emotion cues that map onto both depression and PTSD outcomes.

6 Discussion

6.1 Why Emotion Probes Win

Probes vs. supervised fine-tuning. Fine-tuning BERT updates 110M parameters with only 107 to 163 labeled participants, a feature-to-sample ratio that invites memorization rather than generalization. The result is AUROC 0.77 on DAIC-WOZ that collapses to roughly 0.35 to 0.47 on E-DAIC, where additional ASR noise and a different participant distribution erase whatever surface patterns the model had memorized. Our probes instead fit only a small linear classifier on top of frozen features, which stays stable across both corpora.

Probes vs. pretrained and lexicon features. VADER, Sentence-BERT, and the GoEmotions classifier all avoid the fine-tuning trap by using pretrained representations off the shelf, but they read emotion off surface text rather than the model's internal state. VADER reduces each utterance to a single polarity score, discarding the multidimensional affective signal that distinguishes clinical conditions. This shows up as a gap between its ranking and its decisions: AUROC reaches 0.74 across all three settings, but F1 stays below 0.60 throughout, meaning the polarity signal can rank participants in roughly the right order yet cannot support a clean clinical threshold. Sentence-BERT adds representational richness but still falls behind our probes by 11 points on DAIC-WOZ (0.73 vs. 0.84) and by as much as 25 points on E-DAIC PTSD (0.60 vs. 0.85). The GoEmotions classifier, which shares our 28-emotion vocabulary, is the strongest surface-feature baseline on F1, yet it is 17 points below our probes on DAIC-WOZ AUROC (0.67 vs. 0.84). On E-DAIC, GoEmotions narrows the gap considerably (AUROC 0.80 for depression, 0.81 for PTSD). This is because E-DAIC interviews contain more explicit affective language at the surface, which a text-level emotion classifier can pick up directly. Even so, our probes still lead on both E-DAIC splits with combined calibration (0.87 and 0.85).

Probes vs. direct prompting. Direct prompting asks the same Mistral 24B backbone to output a final yes/no judgment without ever seeing the training data, while our probes turn the model's internal state into continuous scores that a downstream classifier calibrates to the training-set base rate. Our probes beat zero-shot prompting by 13 AUROC points on DAIC-WOZ and 19 points on E-DAIC depression. Few-shot prompting helps on some splits (Mistral one-shot reaches F1 0.698 on E-DAIC depression) but is unstable: the 3-shot setting collapses to F1 0.261 due to context-budget truncation, while our probes give consistent performance across all three settings. Beyond accuracy, direct prompting also returns only a binary label of the prediction, while our probes decompose every prediction into named emotion contributions that align with clinical expectations (Section 6.3).

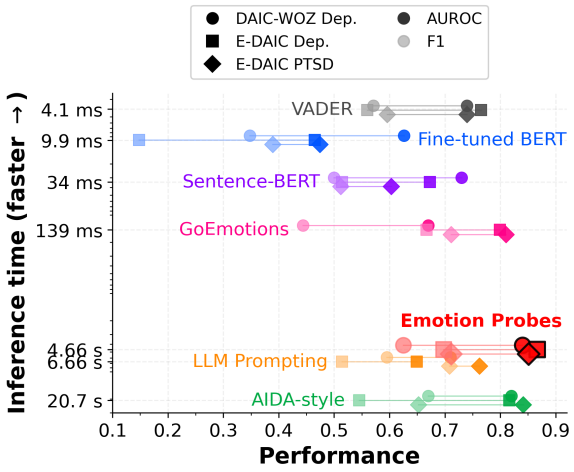

Figure 2: Per-participant inference time vs. AUROC (filled) and F1 (faint) across the three evaluation settings. Marker shape distinguishes DAIC-WOZ Depression (circle), E-DAIC Depression (square), and E-DAIC PTSD (diamond).

Probes vs. clinical-scoring extraction. AIDA-style scoring uses the same backbone and is also extraction-based, but it generates 23 clinical scores autoregressively rather than reading them from a single forward pass. On accuracy the two methods are close, with our probes leading on AUROC across all three settings. The biggest difference is cost: our probes run 4.4× faster than AIDA-style scoring while matching their accuracy, since features are read directly from forward-pass hidden states rather than decoded as text (Section 6.2).

6.2 Computational Cost

Figure 2 plots per-participant inference time against AUROC and F1 for each method and evaluation setting. We measured all methods on a single NVIDIA RTX 2080 Ti GPU, averaging three runs (VADER on CPU). The lightweight baselines (VADER, Sentence-BERT 22M, fine-tuned BERT 110M, GoEmotions 125M) finish in 4 to 139 ms per participant, while the three Mistral 24B methods

Emotion	DAIC-WOZ Depression		E-DAIC Depression		E-DAIC PTSD	
Positive affect						
optimism					−0.25	***
approval	−0.16	*	−0.07	†	−0.24	***
joy					−0.23	***
amusement	−0.21	**				
relief			+0.04	†	−0.18	**
gratitude			+0.06	†	−0.18	**
pride			−0.04	†	−0.18	**
love			+0.03	†	−0.16	**
excitement					−0.14	*
Negative affect						
disgust			+0.24	***		
anger	+0.16	*			+0.03	†
Other						
confusion			+0.15	*		
desire			+0.14	*	−0.04	†
neutral			+0.12	*	−0.08	†

Table 2: Pearson correlations between Lasso-selected emotion features and symptom severity. Each setting uses task-matched probes: depression-calibrated for depression, PTSD-calibrated for PTSD. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; † Lasso-selected but not individually significant (suppressor).

take seconds: 4.66 s for our emotion probes, 6.66 s for LLM prompting, and 20.7 s for AIDA-style scoring. The cost gap among the 24B methods comes from decoding. Our probes generate no output tokens since features are read directly from forward-pass hidden states, while LLM prompting decodes a short binary judgment (∼30 tokens) and AIDA-style scoring decodes around 200 tokens of clinical scores per participant. This makes our probes 1.4× faster than LLM prompting and 4.4× faster than AIDA-style scoring.

6.3 Interpretable Predictions

Our method is interpretable by construction: each of the 28 features is a named emotion, predictions are linear combinations of those features, and the emotion-severity associations can be read off the model directly. We answer two questions: which emotions carry unique predictive content, and how strongly each relates to severity on its own. For the first, we fit LassoCV (10-fold) on the training set and report the selected subset. For the second, we compute Pearson correlations within that subset. Features that survive Lasso but show no individual correlation are suppressors: they improve the model by removing noise from other emotion features rather than tracking severity directly.

The depression probes show a different signal across the two corpora. On DAIC-WOZ, the model highlights amusement and approval as negative correlates and anger as positive, consistent with reduced social engagement and irritability in depression. On E-DAIC, the dominant correlates shift to disgust, confusion, and desire. DAIC-WOZ and E-DAIC surface different depression correlates because the interviews themselves differ in content: DAIC-WOZ's Wizard-of-Oz protocol with general-public participants captures social-affiliative cues (amusement, approval, anger), while E-DAIC's longer, more structured AI-led interviews with a veteran-enriched sample bring out a cognitive-affective signal (disgust, confusion).

The PTSD profile is dominated by reductions in positive emotion. Optimism, approval, joy, relief, gratitude, pride, love, and excitement all correlate negatively with PTSD severity. This is consistent with the emotional numbing and anhedonia symptoms in PTSD, which include diminished interest, detachment, and inability to experience positive emotions. On the same E-DAIC population, PTSD shows broad positive-affect numbing while depression shows targeted increases in disgust and confusion, demonstrating that our probes resolve two clinically overlapping conditions at the level of individual emotions.

6.4 Domain-Specific Calibration

Depression probes are calibrated on depression-related topics, which may not capture PTSD symptoms. We recalibrate the same 28 emotion directions on 50 PTSD-relevant topics spanning the major PTSD symptom clusters (Appendix B), keeping the model and meta-prompt fixed, and additionally construct a *combined* calibration merging all 100 topics. The resulting depression and PTSD probe sets carry distinct geometric structure that aligns with the clinical symptoms of each condition (Appendix D), indicating that changes in calibration content produce real differences in the resulting probe directions. Table 4 reports all three calibrations across both datasets. On DAIC-WOZ, depression calibration is best ($R^2 = 0.426$, AUROC = 0.840), and adding PTSD or combined topics degrades every metric. On E-DAIC, combined calibration performs best for both depression ($R^2 = 0.375$, AUROC = 0.867) and PTSD ($R^2 = 0.529$, AUROC = 0.851).

This contrast follows from how each dataset was collected. DAIC-WOZ contains 189 Wizard-of-Oz sessions with participants recruited from

Task	Calibration	MAE ↓	R^2 ↑	AUROC ↑	F1 ↑
DAIC-WOZ Depression					
	Depression	**3.92**	**0.426**	**0.840**	**0.686**
	PTSD	4.29	0.310	0.810	0.562
	Combined	4.00	0.413	0.831	0.625
E-DAIC Depression					
	Depression	4.57	0.259	0.836	0.619
	PTSD	4.37	0.284	0.858	0.638
	Combined	**4.07**	**0.375**	**0.867**	**0.698**
E-DAIC PTSD					
	Depression	13.15	0.365	0.738	0.714
	PTSD	11.49	0.506	0.844	0.667
	Combined	**11.30**	**0.529**	**0.851**	**0.711**

Table 3: Domain-specific probe calibration across datasets. Depression calibration uses 50 depression-related topics; PTSD uses 50 PTSD topics; Combined merges all 100. Best task result in **bold**.

Craigslist (Gratch et al., 2014), a general-public sample with only PHQ-8 depression labels and few participants likely to have PTSD, so PTSD-calibrated probes extract signal from a content axis that is largely absent. E-DAIC extends the corpus with sessions conducted by a fully autonomous interviewer (Ringeval et al., 2019), expanding recruitment to include veterans with higher trauma exposure, and the resulting test population frequently expresses both depression and trauma. Combined probes capture E-DAIC the best in both depression and PTSD since the interview contains both symptoms. The practical implication is that calibration should be matched to the content coverage of the target corpus, and when multiple symptoms coexist, combined calibration captures richer signal at no additional cost, since adapting to a new condition requires only a new topic list rather than new labels or a retrained model.

7 Conclusion

We presented an emotion probe pipeline for depression and PTSD detection from clinical interviews that reads affective signal from the internal representations of LLM. The pipeline calibrates 28 emotion directions from emotional and neutral message pairs, projects per-turn hidden states onto them, and fits a small linear classifier on the resulting per-participant features. Across DAIC-WOZ depression, E-DAIC depression, and E-DAIC PTSD, our method achieves the highest AUROC among baseline methods (0.846, 0.836, and 0.846), while running 1.4× faster than direct prompting and 4.4× faster than AIDA-style scoring on the same LLM backbone. The Lasso-selected emotion sets are clinically interpretable, and matching the calibration content to the target corpus improves performance with no additional labels or retraining. Together, these results point to emotion probes as a practical and interpretable foundation for AI-driven mental health screening.

8 Limitation

Single model. All experiments use Mistral Small 3.2 24B Instruct. The protocol applies in principle to any causal LM with an accessible residual stream, but whether the calibrated directions carry the same signal in other model families is untested.

Dataset and label limitations. DAIC-WOZ and E-DAIC provide only PHQ-8 and PCL-C self-report labels, which are subject to recall bias and stigma, and the relationship to clinician-assigned diagnosis is not measured. The test splits are also small ($n = 47$ for DAIC-WOZ, $n = 56$ for E-DAIC), which limits their use for drawing clinical conclusions.

Vocabulary choice. We report the 28-direction GoEmotions vocabulary. We also tried smaller vocabularies built from PHQ-8 symptom items, both on their own and combined with GoEmotions, and neither did better. One reason may be that GoEmotions already covers the positive-affect side of depression that PHQ-8 items miss, so the PHQ-8 additions mostly duplicated signal already present.

Synthetic-pair calibration. The probe directions come from synthetic emotional/neutral message pairs that Mistral generates, and they are then read off Mistral's own hidden states on those same pairs. This is a closed loop, so the probes may capture how Mistral writes about emotion rather than how depression and PTSD show up in real patients. The probes still work when applied to PTSD after being calibrated only on depression topics, which is some evidence that they capture more than Mistral's own writing style. Stronger tests on independent clinical interviews are still needed.

References

Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720.

Edward B Blanchard, Jacqueline Jones-Alexander, Todd C Buckley, and Catherine A Forneris. 1996. Psychometric properties of the ptsd checklist (pcl). *Behaviour research and therapy*, 34(8):669–673.

Christy A Blevins, Frank W Weathers, Margaret T Davis, Tracy K Witte, and Jessica L Domino. 2015. The posttraumatic stress disorder checklist for dsm-5 (pcl-5): Development and initial psychometric evaluation. *Journal of traumatic stress*, 28(6):489–498.

Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. *arXiv preprint arXiv:2307.00920*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4040–4054.

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.

Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, pages 69–76.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and 1 others. 2014. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pages 3123–3128. Reykjavik.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Gaoussou Youssouf Kebe, Jeffrey M Girard, Einat Liebenthal, Justin Baker, Fernando De la Torre, and Louis-Philippe Morency. 2025. Llamadrs: Prompting large language models for interview-based depression assessment. *arXiv preprint arXiv:2501.03624*.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Jae-Joong Lee, Jihoon Han, and Choong-Wan Woo. 2026. Interpretable depression assessment using a large language model. *PLOS Digital Health*, 5(2):e0001205.

Ming Li, Yusheng Su, Hsiu-Yuan Huang, Jiali Cheng, Xin Hu, Xinmiao Zhang, Huadong Wang, Yujia Qin, Xiaozhi Wang, Kristen A Lindquist, and 1 others. 2024. Language-specific representation of emotion-concept knowledge causally supports emotion inference. *Iscience*, 27(12).

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

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Mark W Miller, Jennifer L Greif, and Alethea A Smith. 2003. Multidimensional personality questionnaire profiles of veterans with traumatic combat exposure: externalizing and internalizing subtypes. *Psychological assessment*, 15(2):205.

Mark W Miller and Patricia A Resick. 2007. Internalizing and externalizing subtypes in female sexual assault survivors: Implications for the understanding of complex ptsd. *Behavior therapy*, 38(1):58–71.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376.

Julia Ohse, Bakir Hadžić, Parvez Mohammed, Nicolina Peperkorn, Michael Danner, Akihiro Yorita, Naoyuki Kubota, Matthias Rätsch, and Youssef Shiban. 2024. Zero-shot strike: Testing the generalisation capabilities of out-of-the-box llm models for depression detection. *Computer Speech & Language*, 88:101663.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, and 1 others. 2019. 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.

Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 3–9.

Gony Rosenman, Talma Hendler, and Lior Wolf. 2024. Llm questionnaire completion for automatic psychiatric assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 403–415.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, and 1 others. 2026. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*.

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120.

David Watson, Lee Anna Clark, and Greg Carey. 1988. Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of abnormal psychology*, 97(3):346.

Wen Wu, Mengyue Wu, and Kai Yu. 2022. Climate and weather: Inspecting depression detection via emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6262–6266. IEEE.

Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews. *arXiv preprint arXiv:2006.08336*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500.

Xu Zhang, Chenlong Li, Weisi Chen, Jiaxin Zheng, and Feihong Li. 2025. Optimizing depression detection in clinical doctor-patient interviews using a multi-instance learning framework. *Scientific Reports*, 15(1):6637.

Xianbing Zhao, Yiqing Lyu, Di Wang, and Buzhou Tang. 2025a. Predicting depression in screening interviews from interactive multi-theme collaboration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23025–23035.

Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C Oliveira, Robert Desimone, Dominic Dwyer, and Zongyuan Ge. 2025b. It hears, it sees too: Multimodal llm for depression detection by integrating visual understanding into audio language models. *arXiv preprint arXiv:2511.19877*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Implementation Details

Backbone model. Mistral Small 3.2 24B Instruct is loaded with 4-bit NF4 quantization (double quantization, float16 compute). The model has 40 transformer layers and a 5120-dimensional residual stream. Hidden states are extracted at four candidate layers computed as $\{50\%, 60\%, 70\%, 80\%\}$ of model depth, i.e., layers $\{20, 24, 28, 32\}$. Interview transcripts are chunked at 2048 tokens.

Stage 1: Probe calibration. For each of the 28 GoEmotions categories, we generate 50 paired emotional/neutral first-person synthetic prompts spanning diverse topics, using a meta-prompt. Each prompt is passed through the frozen backbone, and the hidden state at the final user-turn token is recorded. For each candidate layer, the emotion direction v_i is computed as the mean difference between the emotional and neutral hidden states and then unit-normalized. To select the layer, we evaluate circumplex alignment: a PCA is fit on the 28 direction vectors, and point-biserial correlations are computed between PC1 and valence polarity (positive/negative) and between PC2 and arousal level (high/low). The layer quality score is $s(l) = 0.6 \cdot |r_{\text{valence}}| + 0.4 \cdot |r_{\text{arousal}}|$, and the layer with the highest score is retained. In practice this is layer 28 (70% depth), and all downstream computation uses only the 28 directions $\{v_1, \ldots, v_{28}\}$ from this layer.

Stage 2: Per-participant features. For each participant p, the clinical interview transcript is tokenized and passed through the same Mistral backbone in 2048-token chunks. Hidden states at the selected layer are collected at each participant turn, and every per-turn hidden state is projected onto each of the 28 emotion directions via inner product, yielding a 28-dimensional emotion vector for that turn. The per-turn vectors are then aggregated across the interview to form the participant-level feature $x_p \in \mathbb{R}^{28}$.

Stage 3: Feature selection. An L1-regularized regression (Lasso) is fit on the training split to select an informative subset $S \subseteq \{1, \ldots, 28\}$ of emotion features. We use scikit-learn's `LassoCV` with 5-fold cross-validation, searching the penalty λ over 100 values spaced logarithmically in $[10^{-4}, 10^{2}]$. A coefficient is treated as nonzero when its absolute value exceeds 10^{-8}. The filter is fit on the training split only, and the surviving directions in S are then passed to the downstream classifiers.

Stage 4: Downstream prediction. The same downstream pipeline is applied to all feature-producing methods. Continuous severity scores are produced by Ridge regression, with α selected from $\{0.01, 0.1, 1, 10, 100, 1000\}$ via 5-fold cross-validation on the training set, scored by negative MAE. Binary classification is produced by Logistic regression, with C selected from $\{0.01, 0.1, 1, 10, 100\}$ via 5-fold cross-validation with class-balanced weights, scored by AUROC. Features are standardized using training-set statistics, and any NaN or inf values are replaced with 0.

B Calibration Materials

A. 28-emotion vocabulary based on GoEmotions (Demszky et al., 2020)	
Positive (12)	admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief
Negative (11)	anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness
Neutral (5)	confusion, curiosity, realization, surprise, neutral
B. 50 depression-relevant topics (Kroenke et al., 2009)	
Daily routines (5)	typical day at home alone; weekend with no plans; cooking and eating alone; scrolling social media; moving to a new city
Social relationships (10)	friends drifting apart; feeling left out; difficulty making adult friends; breakup and losing mutual friends; surface-level conversations; family estrangement; only single one in friend group; reconnecting with an old friend; online vs in-person friendships; conflict with a close friend
Work and purpose (5)	remote work isolation; feeling invisible at work; retirement and loss of structure; job loss and identity; meaningful work achievement
Health and self-care (5)	chronic illness limiting social activity; sleep problems from loneliness; starting an exercise routine; mental health struggles; caring for an aging parent alone
Existential (10)	questioning meaning of life; nostalgia for past relationships; fear of being alone forever; comparing life to expectations; grieving a loss; birthday alone; holiday loneliness; feeling like a burden; self-worth and belonging; finding peace with solitude
Growth (10)	kind interaction with a stranger; joining a community group; adopting a pet; small step toward connection; new hobby; volunteering; therapy progress; gratitude for the chatbot; setting a social goal; feeling hopeful
Specific (5)	seeing couples in public; alone in a crowd; receiving bad news with no one to tell; neighborhood detachment; digital fatigue and craving real connection
C. 50 PTSD-relevant topics (Blevins et al., 2015)	
Intrusion (9)	intrusive memories during daily activities; recurring nightmares; flashback triggered by a sound; emotional reaction to a reminder; physical sensations when remembering; anniversary of the event; triggered by a similar location; vivid images that will not go away; triggered by a specific smell
Avoidance (7)	avoiding conversations about the past; refusing certain places; emotional numbing; avoiding news and media; canceling plans to stay home; not wanting to talk about feelings; changing routines to avoid triggers
Negative mood (10)	self-blame for what happened; world feels completely dangerous; loss of interest in prior activities; detached from loved ones; persistent shame; believing no future exists; survivor guilt; distorted memory of the event; feeling permanently changed; difficulty feeling positive emotions
Arousal (8)	hypervigilance in public; exaggerated startle response; irritability and angry outbursts; severe difficulty concentrating; chronic sleep disturbance; reckless behavior; physical tension and muscle pain; difficulty in enclosed spaces
Growth (8)	first good night of sleep in weeks; opening up about the trauma; therapy progress; returning to a previously avoided place; moment of genuine happiness; helping someone who went through it; recognizing a trigger and coping; rebuilding trust
Specific (8)	being asked about visible scars; encountering the person who caused the trauma; legal proceedings; explaining PTSD to family; impact on intimate relationships; workplace challenges; medication side effects; triggered by a news story

Table 4: Calibration materials. Each topic in B and C is paired with the 28 emotions in A to produce one emotional and one matched neutral first-person message.

Table 4 lists the 28-emotion vocabulary (A) and the 50 depression-relevant (B) and 50 PTSD-relevant (C) topics used for probe calibration. For each (emotion, topic) pair we generate one emotional and one matched neutral first-person message ($50 \times 28 \times 2 = 2{,}800$ per outcome); per-turn activations are averaged within each variant, and the mean-difference direction in Eq. 1 gives the calibrated probe.

C Token-Level Activation Examples

Figure 3: Per-token projection $\langle h_t, v_j \rangle$ onto three emotion directions (joy, anger, neutral) for two example scenarios, one from a non-depressed participant and one from a depressed participant.

The participant-level feature x_p in the main pipeline is built from per-turn projections $\langle h, v_j \rangle$, where h is the hidden state at the final token of each participant turn. To make the projection operation concrete and to show what the emotion probes actually respond to, Figure 3 visualizes the same operation at the finer token level, $\langle h_t, v_j \rangle$, on two example scenarios: one from a non-depressed participant and one from a depressed participant. We show three of the 28 emotion directions, chosen to cover the main affective axes: joy, anger, and neutral. The token-level view is for inspection only and is not part of the inference pipeline.

The three directions diverge across the two passages in the way one would predict from the depression literature. Joy (blue) is broadly active on the non-depressed passage, including on tokens that are not overtly positive ("starting to plan a trip"), but stays consistently faint on the depressed passage. Anger (red) shows the opposite pattern: it is weak on the non-depressed passage and strongest on rumination clauses such as "I keep" and "and I can't" on the depressed passage. Neutral (green) activates on common function words (articles, helping verbs, "I", punctuation) in both passages with no clear difference between them, serving as a sanity check. The per-token activations shown here are aggregated to per-turn and then per-participant features in the main pipeline.

D Depression vs. PTSD Probes

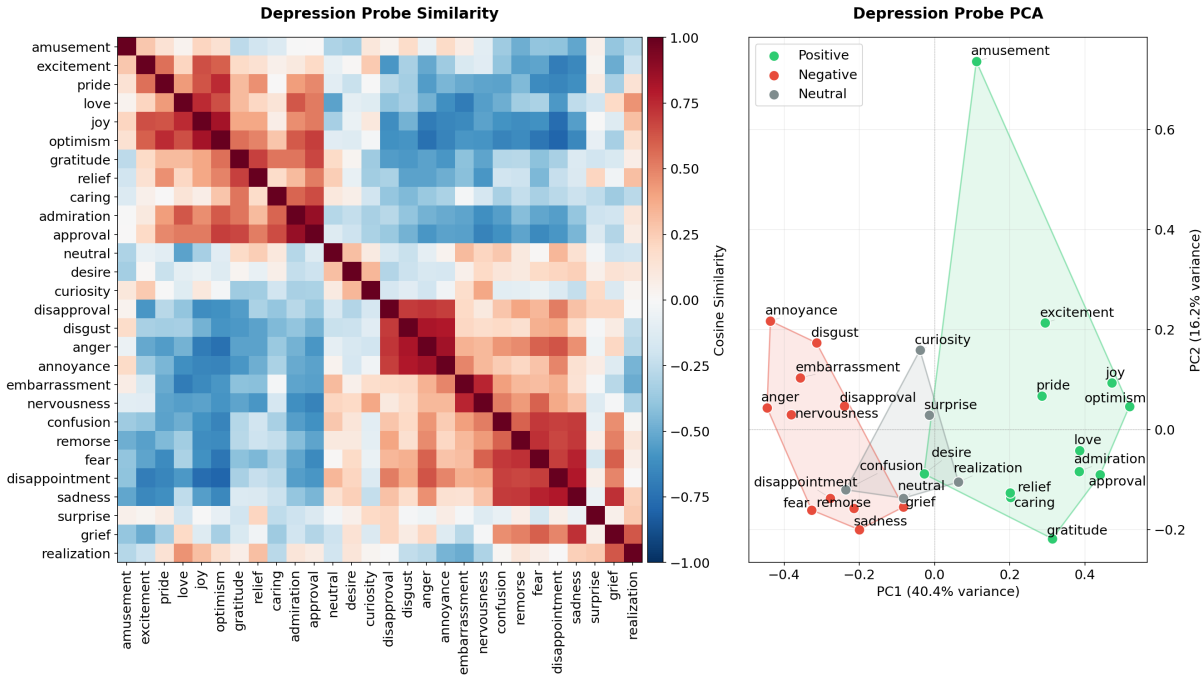

Figure 4: Depression probes. Left: pairwise cosine similarity of the 28 residual probes. Right: PC1–PC2 projection over positive, negative, and neutral valence groups.

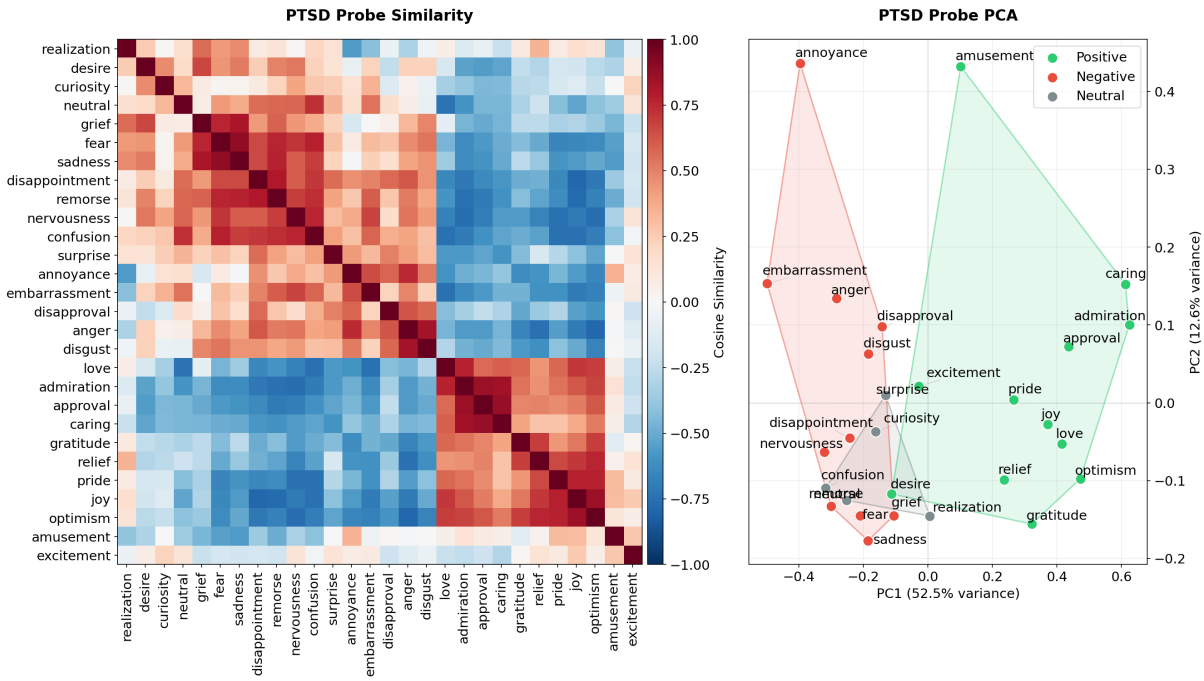

Figure 5: PTSD probes. Left: pairwise cosine similarity of the 28 residual probes. Right: PC1–PC2 projection over positive, negative, and neutral valence groups.

We compare the structure of the 28 emotion probes in depression and PTSD through pairwise cosine similarity and a PC1–PC2 projection (Figures 4 and 5). All figures use residual probes, which subtract the mean of the 28 directions from each probe and normalize the result to expose emotion-specific structure, following Marks and Tegmark (2023). The cosine similarity matrices (left of each figure) reorder rows

and columns by hierarchical clustering so that probes pointing in similar directions appear adjacent. The PCA projections (right of each figure) compress the 28 probes into two dimensions by finding the two directions along which the probes vary the most, where the percentage of variance explained by each axis indicates how much of the original structure is preserved in the projection. If the probes encode psychologically meaningful structure, these two axes are expected to align with valence and arousal, the two primary dimensions of affect.

Depression vs. PTSD probes. We compare the depression and PTSD probe geometries in Figures 4 and 5. Both probe sets separate positive and negative valence along PC1, but the finer structure differs in clinically meaningful ways. For the depression probes (Figure 4), PC1 (40.4%) cleanly splits positive from negative emotions, and the negative emotions show narrower spread along PC2 (16.2%), forming a single compact group rather than dividing into smaller clusters. This is consistent with how depression often presents through emotional numbness, blunted reactivity, and withdrawal rather than through sharply distinguished negative emotions. For the PTSD probes (Figure 5), PC1 (52.5%) again captures valence, but PC2 (12.6%) now separates the negative emotions into two clusters visible in both the heatmap and the PCA: a lower cluster centered on grief, fear, sadness, and nervousness, and an upper cluster centered on annoyance, embarrassment, anger, and disgust. This split mirrors how PTSD's negative symptoms are organized into two distinct clusters: a distress-and-fear cluster (grief, sadness, anxiety) and an anger-and-reactivity cluster (irritability, hostility, angry outbursts) (Miller et al., 2003; Miller and Resick, 2007). These differences confirm that domain-specific probes capture different emotion structures.

Probe redundancy and multicollinearity. The cosine similarity heatmaps show clear off-diagonal structure in both probe sets, with several tight clusters of positive emotions and several tight clusters of negative emotions reaching pairwise cosines above 0.7. This redundancy is expected since the GoEmotions taxonomy was built for fine-grained text classification rather than for independent affective dimensions, so neighboring emotion categories share much of their semantic content. When the 28 features are passed into a downstream regression on only 107 to 163 training participants, this redundancy turns into multicollinearity that inflates the risk of overfitting. Therefore, feature selection is needed to retain a smaller subset of directions that captures most of the predictive signal without the redundancy.

E Transcription Validation (E-DAIC)

Pipeline. E-DAIC participant audio is transcribed with ElevenLabs Scribe using speaker diarization. The participant is identified as the diarized speaker with more total words per session, since interviewer bleed-through into the participant channel is sparse.

Comparison setup. To validate transcription quality against the DAIC-WOZ gold transcripts on the 171 overlapping sessions, we filter to participant turns on both sides, drop the `xxx` unrecognizable-speech markers and angle-bracket annotations on the gold side, concatenate participant text in time order, and normalize both sides with the Whisper English text normalizer before scoring with `jiwer`.

Exclusions. We exclude 9 sessions flagged for data-quality issues in the DAIC-WOZ documentation, and 9 additional sessions with per-session WER above 25%. Manual inspection shows the latter are diarization-outlier sessions where the participant and interviewer channels were swapped or merged, not ASR-quality failures.

Results. On the remaining 171 sessions, corpus-level WER is 5.60% (mean per-session 6.05%, median 4.98%). The error mix is dominated by insertions (2.76%), with substitution rate 1.75% and deletion rate 1.09%, consistent with Scribe transcribing filler words and false starts that human transcribers omit.

F Baseline Implementation Details

This appendix documents the implementation of each baseline summarized in Section 4.3. All feature-producing baselines share the same Ridge and Logistic pipeline as our probes (Stage 4); only the feature extraction differs.

Lexicon/pretrained features. VADER (Hutto and Gilbert, 2014) returns four polarity scores per utterance (positive, negative, neutral, compound), averaged per participant. Sentence-BERT (Reimers and Gurevych, 2019) uses `all-MiniLM-L6-v2` with mean pooling to encode each utterance into a 384-dimensional embedding, also averaged per participant.

GoEmotions classifier. GoEmotion (Liu et al., 2019; Demszky et al., 2020) is a RoBERTa model fine-tuned on Reddit comments labeled with the 28 GoEmotions categories. It produces a 28-dimensional probability distribution per utterance, averaged per participant.

Fine-tuned BERT. We fine-tune BERT (Devlin et al., 2019) end-to-end with class-weighted cross-entropy and AdamW, using learning rate 2×10^{-5}, batch size 8, and up to 10 epochs with early stopping on dev AUROC. The direct-classification head reports binary predictions, and the [CLS]-feature variant extracts the 768-dimensional penultimate-layer hidden state for the standard Ridge and Logistic pipeline.

AIDA-style inference. We reproduce Lee et al. (2026), originally run on LLaMA 3.1-70B, using Mistral Small 3.2 24B Instruct and GPT-4o as the scoring models. The LLM rates the interview on 23 clinically grounded dimensions (8 from PHQ-8 and DSM-5 criteria, 5 linguistic patterns, 10 cognitive distortions) in a single structured generation, each on a 0 to 10 scale. Dialogues are truncated to 6000 characters, keeping the first and last 3000 characters; unparsed dimensions default to 5.0 (default rate 8.8% on DAIC-WOZ). The resulting 23-dimensional vector feeds the standard Ridge and Logistic pipeline.

Direct LLM prompting. The LLM reads a Participant/Ellie dialogue with standard severity ranges and threshold definitions, and returns a single binary judgment in zero-shot, 1-shot, and 3-shot variants with examples balanced across classes. Both GPT-4o (temperature 0) and Mistral Small 3.2 24B (greedy decoding) are evaluated. The dialogue is capped at 8000 characters, reduced proportionally when exemplars are present. For Mistral, AUROC is computed from hard binary predictions, equivalent to balanced accuracy (Yang et al., 2024).