
EmoTrack: Clinical-Semantic Modeling for Text-Based Depression Severity Estimation

Zhaomin Wu, Jiayi Li, Bingsheng He

Department of Computer Science

National University of Singapore

zhaomin@nus.edu.sg, li.jiayi@u.nus.edu, dcsheb@nus.edu.sg

Abstract

Text-based counseling provides a valuable source of information for assessing depression severity. We study prediction of the total score on the eight-item Patient Health Questionnaire (PHQ-8), a self-report measure of depression severity, from counseling transcripts. Clinical-based methods rely mainly on large language model (LLM) inference to obtain structured session-level assessments, but these assessments provide limited information about which utterances support each score. Training-based methods train predictors directly on sentences or their semantic embeddings and preserve local conversational detail, but must learn clinical structure from limited labeled transcripts. Integrating a small set of clinical feature scores with a long sequence of high-dimensional utterance representations is challenging, as simple fusion can overemphasize dialogue evidence and fail to link clinical features to supporting utterances. Combining structured clinical assessments with sentence-level semantics, we propose EMOTRACK, which jointly encodes clinical feature scores and utterance representations to predict the total PHQ-8 score. For longitudinal assessment across successive sessions, the model can also incorporate a compressed representation of the preceding session as optional memory. On the real-world DAIC-WOZ benchmark, EMOTRACK reduces mean absolute error from 2.8234 to 2.4708 relative to the strongest evaluated baseline. Across five settings covering single-session estimation, cross-dataset transfer without adaptation, and longitudinal assessment, it reduces normalized aggregate error by approximately 6.1% relative to the strongest baseline.

1 Introduction

Online mental health services, including text-based counseling, offer an accessible route to timely support for people with mental health conditions [8]. Monitoring depression severity is important for understanding clients’ needs within these services. The Patient Health Questionnaire (PHQ) is among the most widely used instruments for depression assessment [11]. Its eight-item version (PHQ-8) is a standardized self-report measure of depressive symptom severity, with a total score ranging from 0 to 24 [14]. We study prediction of this score directly from a counseling-session transcript.

Existing methods fall largely into two categories. *Clinical-based methods* rely mainly on LLM inference to derive a session-level assessment by completing a questionnaire or scoring clinician-style features [17, 34]. These assessments supply explicit clinical structure through predefined features and symptom categories, but provide limited information about the contribution of individual turns. *Training-based methods* train predictors directly on sentences or on utterance embeddings that represent their semantic content as vectors [2, 16, 25, 32]. They preserve local conversational detail, but must learn clinically relevant structure from scarce and costly transcript annotations. Combining clinical evidence from LLM-extracted feature scores with conversational evidence expressed in individual utterances can bring clinical structure into transcript modeling under limited supervision.

This combination presents two challenges for single-session estimation and a further challenge for longitudinal use. **(i) Heterogeneous evidence must interact while preserving the meaning of clinical scores.** Clinical assessment produces a small set of named scores, whereas dialogue semantics forms a variable-length sequence of utterance vectors. Combining these representations requires retaining what each clinical score means while allowing it to interact with individual utterances. **(ii) Different PHQ-8 items may require different evidence.** Item-specific evidence may appear in different parts of the conversation: a turn discussing insomnia may inform the sleep item, while a statement about feeling worthless may inform the low-self-worth item. With a pooled readout that predicts all scores from one compressed session representation, each item must rely on the same summary of the available evidence. **(iii) Historical context may be unavailable or outdated.** A preceding session can clarify an indirect statement when the client’s condition persists, but can become misleading after a change in symptoms and is unavailable for first or isolated sessions.

Our key idea is to represent heterogeneous clinical assessments and client turns as distinct evidence that can be selectively retrieved for each PHQ-8 item rather than globally pooled. EMOTRACK implements this idea with three corresponding components. **(i)** The model represents clinical feature scores and client turns as tokens, fixed-dimensional vectors in a common representation space. A clinical token combines a vector representation of the feature score with embeddings that identify the feature and its clinical group. A client-turn token is obtained by projecting the turn’s semantic embedding into the same representation space. This preserves the identity of each clinical feature and the semantic content of each client turn as separate model inputs. **(ii)** The model uses one learned query vector per PHQ-8 item to weight the encoded evidence for total-score estimation. **(iii)** For longitudinal use, EMOTRACK treats history as optional auxiliary evidence: it compresses client-turn representations from the preceding session into a small memory, retrieves information from this memory separately for each PHQ-8 item, and gates the retrieved information before adding it to the current-session representation. Current-session evidence remains the basis for prediction, while learned gates control how much historical information each item receives.

This paper makes three main contributions. (1) We introduce EMOTRACK, a model that jointly encodes prompted clinical evidence and utterance-level semantics and estimates the PHQ-8 total score. (2) We extend this formulation with optional previous-session memory while retaining the same current-session architecture when history is absent. (3) We evaluate the model on the real clinical interview benchmark DAIC-WOZ, on transfer from DAIC-WOZ to E-DAIC without adaptation, and on the three synthetic longitudinal benchmarks of LongCounsel-8, with controlled readout, predictor, memory, objective, and representation ablations. On DAIC-WOZ, EMOTRACK lowers mean absolute error in total-score prediction from 2.8234 to 2.4708. Across five evaluation settings, it reduces normalized aggregate error by approximately 6.1% relative to the strongest baseline.

2 Related Work

We consider methods for transcript-based depression severity prediction. These methods obtain clinical information through prompted inference or supervised learning.

Clinical-based methods. Clinical-based methods rely mainly on inference and produce session-level assessments with limited attribution to individual turns. They provide diagnostic criteria to an LLM and derive a severity estimate by completing a questionnaire or rating clinician-style features that are subsequently mapped to a score [17, 24, 34]. Prompted criteria reduce the need for labeled training data and provide interpretable intermediate outputs. Their session-level outputs do not explicitly identify which turns support each assessment, limiting the evidence available to downstream predictors. EMOTRACK retains the resulting clinical scores but represents them jointly with client turns, enabling the readout to relate each score to supporting utterances and optional historical context.

Training-based methods. Training-based methods are limited by data scarcity because annotated counseling transcripts are small and costly to obtain. These methods train predictors directly on sentences or their semantic embeddings, using architectures such as RoBERTa or sentence-embedding encoders for PHQ-8 prediction [16], hierarchical attention over utterances for PHQ-8 item prediction [25], and retrieval architectures adapted from questionnaire-score prediction [32]. Utterance-level representations preserve local evidence, but the model must learn clinical structure from limited

Table 1: Methods for transcript-based depression severity prediction. *Clinical knowledge*: criteria supplied during inference rather than learned from labeled transcripts; *Training*: trained on sentence or sentence embeddings; *Longitudinal*: uses earlier sessions from the same client.

Study	Clinical knowledge	Training	Longitudinal	Target scale
Lau et al. [16]	✗	✓	✗	PHQ-8
Milintsevich et al. [25]	✗	✓	✗	PHQ-8
EnsemBERT [32]	✗	✓	✗	BDI-II
PsychiatricPrism [2]	✗	✓	✗	PHQ-8
LMIQ [34]	✓	✗	✗	PHQ-8
AIDA [17]	✓	✗	✗	PHQ-8
EMOTRACK (ours)	✓	✓	✓	PHQ-8

supervision. EMOTRACK reduces this requirement by obtaining clinical structure from prompted scores while retaining utterance-level semantics in a 1.02M-parameter trainable model.

Existing systems generally treat clinical-based inference and training-based utterance modeling separately (Table 1). Our focus is to jointly represent prompted clinical features and individual client turns, with a separate query for each PHQ-8 item and an optional extension for previous-session context. Appendices A.1 and A.2 provide a more detailed comparison of methods and datasets.

3 Problem Formulation

We estimate depression severity from a counseling *session*, a complete encounter between a client and a counselor. Its transcript records a sequence of *turns*, each an individual utterance by either participant. The PHQ-8 assesses depressive symptoms through eight questionnaire *items*; our prediction target is its total score, a scalar value in $[0, 24]$ [14].

Single-session estimation. Given the transcript $\mathcal{T}^{(t)}$ of session t , the predictor returns a single scalar estimate $\hat{Y}^{(t)} = f_{\theta}(\mathcal{T}^{(t)}) \in [0, 24]$ of the reference total $Y^{(t)}$. Evaluation measures the error between the predicted and reference total scores.

Extension to multiple sessions. For repeated sessions, the available history is $\mathcal{H}^{(t-1)} = \{\mathcal{T}^{(1)}, \dots, \mathcal{T}^{(t-1)}\}$, and the prediction becomes $\hat{Y}^{(t)} = f_{\theta}(\mathcal{T}^{(t)}, \mathcal{H}^{(t-1)}) \in [0, 24]$. The default EMOTRACK model uses the immediately preceding session $\mathcal{T}^{(t-1)}$ as optional context; Appendix F.6 examines longer histories. When $\mathcal{H}^{(t-1)} = \emptyset$, the model uses only the current transcript.

4 Approach: EMOTRACK

Design rationale. EMOTRACK estimates total PHQ-8 severity from a counseling *session*, a complete client–counselor encounter (Section 3; Figure 1). The encoder preserves clinical-feature identities while jointly encoding their scores and client *turns*, the individual utterances contributed by the client. A separate query for each PHQ-8 *item*, a questionnaire question about a depressive symptom, weights evidence for total-score estimation. Optional memory adds context from the preceding session when available. Appendix B provides the forward algorithm.

4.1 Input Representation

Given the current-session transcript $\mathcal{T}^{(t)}$, EMOTRACK constructs one token sequence from structured clinical features and another from client turns. The dialogue branch uses client turns, following the speaker-selection comparison of client, counselor, and combined inputs in Section 5.3. The clinical branch adopts the existing AIDA schema [17]: an LLM rates $F=23$ clinician-style features in three groups (8 PHQ-aligned symptom indicators, 5 linguistic-pattern features, and 10 cognitive-distortion markers) on a $[0, 10]$ scale. We z-normalize the scores $\{z_i\}_{i=1}^F$ using training-split statistics. The input-source ablation in Section 5.3 evaluates the two branches.

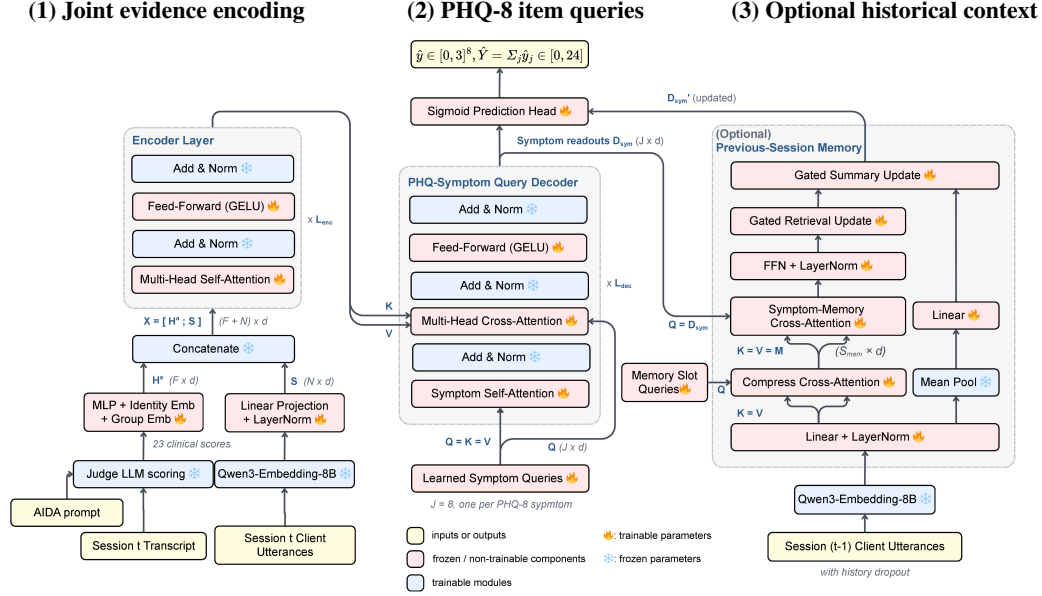


Figure 1: EMOTRACK overview. (1) Joint encoding integrates clinical features and dialogue semantics. (2) PHQ-8 item queries organize evidence for total-score estimation. (3) Optional memory incorporates previous-session context.

Clinical-feature branch. A scalar score alone does not identify the clinical feature it measures. We therefore encode the value and its identity separately. For each normalized feature value z_i , we construct a token $\mathbf{h}_i = \text{MLP}_{\text{score}}(z_i) + \mathbf{e}_i^{\text{feat}} + \mathbf{e}_{g(i)}^{\text{group}}$, where the value projection $\text{MLP}_{\text{score}}(\cdot)$ maps the scalar feature score into the model space, $\mathbf{e}_i^{\text{feat}} \in \mathbb{R}^d$ is a learned embedding identifying the i -th feature, and $\mathbf{e}_{g(i)}^{\text{group}} \in \mathbb{R}^d$ is a learned embedding indicating the clinical group of that feature. Here, $g(i) \in \{1, 2, 3\}$ maps each feature to one of the three groups defined above. The feature and group embeddings keep clinical inputs distinguishable even when their numerical scores are equal. Collecting all feature tokens gives the structured clinical representation $\mathbf{H}^c = [\mathbf{h}_1, \dots, \mathbf{h}_F] \in \mathbb{R}^{F \times d}$.

Dialogue branch. A frozen sentence encoder, whose parameters remain fixed during predictor training, maps each client turn to an utterance embedding $\mathbf{u}_n \in \mathbb{R}^{d_e}$ for turn n . We project this vector as $\mathbf{s}_n = \text{LayerNorm}(W_{\text{proj}}\mathbf{u}_n + \mathbf{b}_{\text{proj}})$. Counselor turns and empty utterances are removed first; sessions with more than N_{max} remaining client turns are truncated by keeping the first N_{max} turns in chronological order, and each utterance is encoded as a single fixed-size embedding regardless of length. We use $N_{\text{max}}=80$ for DAIC-WOZ; the LongCounsel-8 sessions contain 20 client turns each and are not truncated. This gives $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_N] \in \mathbb{R}^{N \times d}$. Appendix F.4 reports the context-length ablation.

4.2 Encoder-Decoder Backbone

The encoder-decoder backbone takes the clinical tokens $\mathbf{H}^c$ and client-turn tokens $\mathbf{S}$ as input and produces one learned representation for each PHQ-8 item.

We first concatenate the two token sequences into a single current-session token sequence, $\mathbf{X} = [\mathbf{H}^c; \mathbf{S}] \in \mathbb{R}^{(F+N) \times d}$. Each clinical feature and each client turn remains a separate token; concatenation here joins token sequences without pooling either source. A padding mask makes self-attention ignore padded client turns. The concatenated sequence is then processed by Transformer encoder layers. In each encoder layer, $\mathbf{X}$ is projected into queries, keys, and values, and self-attention updates each token by attending to all other clinical and client-turn tokens in the current session. This yields a contextualized representation in which each token aggregates relevant evidence from both branches. We use $L_{\text{enc}}=2$ encoder layers on the longitudinal datasets and $L_{\text{enc}}=3$ on DAIC-WOZ, with $L_{\text{dec}}=4$ decoder layers in both cases; this moderate depth is competitive with both shallower and deeper

variants (Appendix F.2). Replacing the EMOTRACK predictor with standard regressors over the same features substantially worsens PHQ-8 prediction, supporting the attention-based backbone (Table 3).

PHQ-8 item query readout. EMOTRACK uses $J=8$ learned query vectors, one for each PHQ-8 item, collected in $\mathbf{Q} \in \mathbb{R}^{J \times d}$. In the decoder, each query cross-attends to the encoded tokens, computing a weighted combination of the evidence represented by those tokens. This allows the sleep and low-self-worth items, for example, to use different combinations of clinical features and utterances from the same session. The decoder returns item representations $\mathbf{D} \in \mathbb{R}^{J \times d}$, where each row corresponds to one PHQ-8 item. Retaining a separate representation for each questionnaire item provides an architectural prior that organizes total-score estimation around the PHQ-8 structure. The readout ablation evaluates its effect on total-score accuracy (Table 2). Appendix G.2 examines the intermediate item predictions.

4.3 Training Objective

The backbone produces item-indexed representations $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_J]^\top \in \mathbb{R}^{J \times d}$. A linear-plus-sigmoid head maps each $\mathbf{d}_j$ to an intermediate item score $\hat{y}_j = 3 \cdot \sigma(\mathbf{w}_j^\top \mathbf{d}_j + b_j) \in [0, 3]$, and the predicted total PHQ-8 severity is $\hat{Y} = \sum_{j=1}^J \hat{y}_j$. We supervise these predictions at two granularities: a Huber loss on the aggregate severity, $\mathcal{L}_{\text{agg}} = \text{Huber}(\hat{Y}, Y)$ with $Y = \sum_j y_j$, and an auxiliary item-level Huber loss $\mathcal{L}_{\text{sym}} = \frac{1}{J} \sum_j \text{Huber}(\hat{y}_j, y_j)$ that encourages each intermediate prediction to match its corresponding item label, following symptom-resolved depression estimation [2, 25]. The full objective combines them with a scalar weight $\lambda_{\text{sym}} \geq 0$,

$$\mathcal{L} = \mathcal{L}_{\text{agg}} + \lambda_{\text{sym}} \mathcal{L}_{\text{sym}}, \quad (1)$$

where we use $\lambda_{\text{sym}}=0.5$ on LC8-QWEN35-35B and $\lambda_{\text{sym}}=0$ on DAIC-WOZ and LC8-GPT54-MINI; $\lambda_{\text{sym}}=0$ recovers aggregate-only supervision (Section 5.3). The total score $\hat{Y}$ is the reported target; $\mathcal{L}_{\text{sym}}$ acts as a regularizer on the readout. We optimize with AdamW [21], using a warmup-decay learning-rate schedule, gradient clipping, and early stopping on validation performance.

Optional false-negative penalty. Mean absolute error weights over- and under-prediction equally, while a screening application may place greater weight on missed positives. At a PHQ-8 threshold of 10, sessions with reference scores of at least 10 are screen-positive, indicating moderate or higher severity [14]. To penalize predictions below this threshold for such sessions, EMOTRACK supports an optional loss term,

$$\mathcal{L}_{\text{FN}} = \lambda_{\text{FN}} \mathbf{1}[Y \geq 10] \max(0, 10 - \hat{Y}), \quad (2)$$

which is added to Eq. (1). The term is active only when a screen-positive session is predicted below the threshold, targets false negatives, and contributes zero direct loss on the remaining sessions; $\lambda_{\text{FN}}=0$ recovers the default objective, and Appendix G.1 reports the trade-off that λ_{FN} controls.

4.4 Optional Previous-Session Memory

Previous-session context can provide continuity when symptoms persist, but may be outdated after a change in the client’s condition. EMOTRACK therefore uses it as optional auxiliary evidence, with the current-session item representations serving as the basis for prediction. Through its *retrieval path*, the memory module compresses prior utterances into a small set of vectors that each item can attend to; its *summary path* supplies an averaged representation of the preceding session. Learned gates scale both updates before adding them to the current-session item representations. Compression limits the amount of historical information passed to the predictor, while the gates learn how much to incorporate. The default model uses the immediately preceding session, and Appendix F.6 examines longer histories.

Let $\{\mathbf{u}_m^{(t-1)}\}_{m=1}^M$ denote the client-turn embeddings from session $t-1$. A dedicated projection maps them into the model space, yielding previous-session client tokens $\mathbf{V}^{(t-1)} \in \mathbb{R}^{M \times d}$.

Retrieval path. Learned slot queries $\mathbf{P} \in \mathbb{R}^{S \times d}$ attend to the projected previous-session client tokens, compressing them into S memory slots that summarize the preceding session:

$$\mathbf{M} = \text{LN}(\text{MHA}(\mathbf{P}, \mathbf{V}^{(t-1)}, \mathbf{V}^{(t-1)})) \in \mathbb{R}^{S \times d}, \quad (3)$$

where $\text{MHA}(\cdot)$ denotes multi-head attention and $\text{LN}(\cdot)$ denotes layer normalization. Each current-session item representation in $\mathbf{D}$ then attends to this compressed memory, yielding a retrieved history representation $\mathbf{C} \in \mathbb{R}^{J \times d}$. We incorporate the retrieved history through a gated residual update that scales its contribution before adding it to the current-session representation, $\mathbf{D} \leftarrow \mathbf{D} + \sigma(W_g[\mathbf{D}; \mathbf{C}]) \odot W_o \mathbf{C}$. The sigmoid gate produces element-wise weights between zero and one. The weights depend on both the current item representation and retrieved history, allowing the contribution to vary across items and representation dimensions. The combined slot-based retrieval and summary path is evaluated in Section 5.3; Appendix F.5 reports the memory-capacity ablation.

Summary path. In addition to slot-based retrieval, we construct a coarse summary of the previous session by mean-pooling the projected previous-session client tokens and mapping the result into the model space. Let $\bar{\mathbf{v}} \in \mathbb{R}^d$ denote this summary vector. We broadcast it to all J items and apply a second gated residual update, $\mathbf{D} \leftarrow \mathbf{D} + \sigma(W_{g'}[\mathbf{D}; \bar{\mathbf{v}}]) \odot \bar{\mathbf{v}}$. This provides each item with a coarse global view of the preceding session. After these refinements, the updated item representations are passed to the prediction head described in Section 4.3.

History dropout. During training, we apply history dropout by randomly omitting previous-session memory with probability p_{hist} , exposing the predictor to cases with and without history. The module is omitted in single-session settings. We use $p_{\text{hist}}=0.1$ because light dropout performs competitively in the comparison (Appendix F.3).

5 Experiments

We evaluate how clinical–semantic modeling supports accurate PHQ-8 estimation across single-session, cross-dataset transfer, and longitudinal settings. For selected baselines, we use the evaluation results reported in the LongCounsel-8 benchmark study [19]. Controlled ablations examine the contribution of each input source, the PHQ-8 item query readout, and the optional extensions. Appendix D gives implementation details; Appendices E–G report full results and analyses.

5.1 Experimental Settings

Datasets. We use five evaluation sets. DAIC-WOZ [9] contains 189 real single-session clinical interviews with self-reported PHQ-8 labels; we use its official participant-level split of 107 training, 35 validation, and 47 test interviews. E-DAIC [33] extends the same interview protocol to further participants; we evaluate frozen cross-dataset transfer by applying a model trained and selected on DAIC-WOZ directly to E-DAIC, with no further training or model selection on the target set. The E-DAIC labels are used only for scoring. LongCounsel-8 [19] is a synthetic benchmark of counseling trajectories, each an ordered sequence of five sessions for one client, with session-level PHQ-8 supervision, grounded in real client profiles, depression trajectories, symptom compositions, and counseling patterns. We use LC8-QWEN35-35B, the main set of 3,690 trajectories generated with Qwen3.5-35B-A3B and split into 2,583 training, 369 validation, and 738 test trajectories, and LC8-GPT54-MINI, a 369-trajectory set generated with GPT-5.4-mini. For LC8-GPT56-LUNA, we use a randomly sampled subset of 369 trajectories from the full GPT-5.6 Luna release due to computational constraints. Thus, LC8-GPT54-MINI and the selected LC8-GPT56-LUNA subset are matched in trajectory count, whereas LC8-GPT54-MINI differs from LC8-QWEN35-35B in both generator and scale. Appendix C provides the dataset and split details.

Model inputs and training. The frozen turn encoder is Qwen3-Embedding-8B [45], which produces a 4,096-dimensional vector per utterance. Following AIDA [17], Qwen3.5-35B-A3B [31] extracts the 23 clinical feature scores. Each extraction seed is paired across all downstream methods that use these features. We train EMOTRACK with AdamW [21] (learning rate 10^{-3} , weight decay 10^{-2}), batch size 32, linear warmup over the first 5% of steps followed by cosine decay, gradient clipping at norm 1.0, and early stopping with patience 8. One EMOTRACK seed trains in about 5 minutes on a single GPU. Appendix D lists the hyperparameters and hardware.

Metrics. Mean absolute error (MAE) averages the absolute difference between predicted and reference PHQ-8 totals. Normalized aggregate error (NAE) is $\text{NAE}(m) = \sqrt{\frac{1}{5} \sum_{j=1}^5 (x_{m,j}/U_j)^2}$,

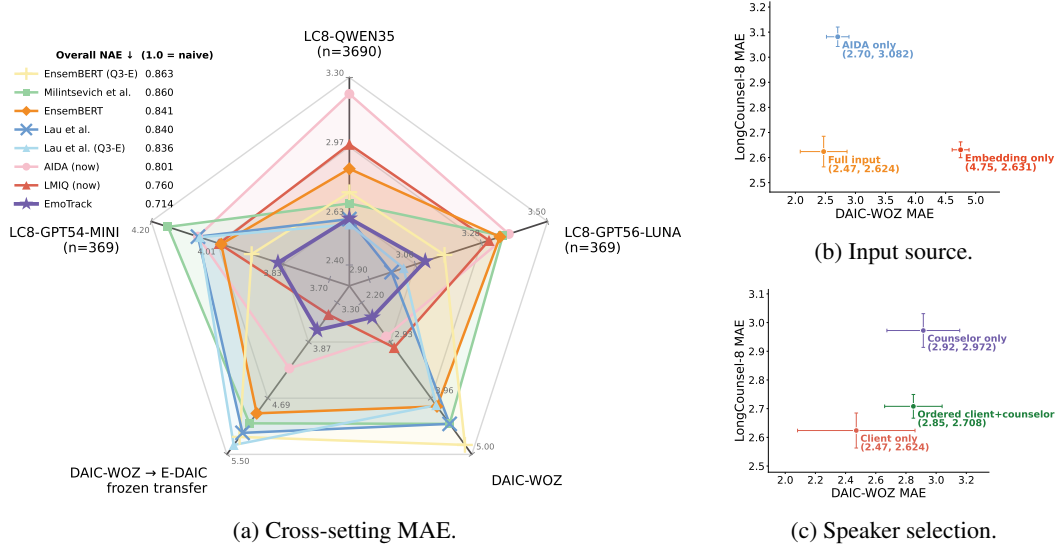


Figure 2: Cross-setting performance comparison and input ablations of EMOTRACK. (a) Axes show raw MAE on separate linear scales; the legend reports NAE, independent of these plotting limits. Q3-E denotes Qwen3-Embedding. (b,c) show MAE for input-source and speaker-selection ablations.

where $x_{m,j}$ is the method’s five-seed mean test MAE in setting j and U_j is the test MAE of a constant predictor that always outputs the median PHQ-8 training label. Each setting receives equal weight; lower values are better, zero denotes perfect prediction, and one denotes the constant-predictor reference. Mean signed error averages predicted minus reference scores. At a PHQ-8 threshold of 10, the false-negative rate (FNR) is the proportion of reference-positive cases predicted negative, and the false-positive rate (FPR) is the proportion of reference-negative cases predicted positive.

Evaluation. Unless otherwise stated, we select model checkpoints and hyperparameter settings by validation MAE. We report the mean and sample standard deviation of test MAE over five extraction/training seeds. For frozen transfer, the constant-predictor reference retains the DAIC-WOZ training median when scoring E-DAIC.

Baselines. The baselines span the two main approaches to transcript-based severity estimation: Clinical-based methods that rely mainly on LLM inference and Training-based methods that train predictors directly on sentences or their semantic embeddings. (1) **AIDA** [17] estimates severity from prompted clinical features. (2) **LMIQ** [34] completes symptom questionnaires from transcripts using an LLM. (3) **Lau et al.** [16] use neural transcript encoding for severity prediction. (4) **Milintsevich et al.** [25] estimate overall severity through hierarchical symptom prediction. (5) **EnsemBERT** [32] predicts severity through symptom-based retrieval. For Lau et al. and EnsemBERT, we evaluate both the original encoders and Qwen3-Embedding variants. For methods that can concatenate history, “now” uses only the current session and “all” uses available prior sessions. We select EnsemBERT’s hyperparameters on the DAIC-WOZ validation split and use the selected setting throughout (Appendix E.3). All comparisons share dataset protocols and extraction seeds.

5.2 Main Results

Performance across settings. Figure 2a shows that EMOTRACK achieves the **lowest normalized aggregate error across all five settings**, combining strong single-session accuracy with transfer and longitudinal performance. Its NAE is 0.714, compared with 0.760 for LMIQ (now), the strongest baseline, an approximately 6.1% reduction computed from the displayed scores. The aggregate result supports joint clinical–semantic modeling as an effective approach across evaluation regimes.

Real single-session evaluation. Figure 2a shows that EMOTRACK has the **lowest mean MAE on DAIC-WOZ**, supporting its use for transcript-based severity estimation with limited labeled interviews. It reaches 2.4708 ± 0.3892 , compared with 2.8234 ± 0.2099 for AIDA, the strongest

baseline, a 12.5% reduction (Appendix Table 8). The prompted methods are the closest competitors: LMIQ reaches 3.0285, while every trained baseline lies above 4.09. The participant-level paired permutation test against AIDA yields $p = 0.1901$ on the 47-participant test set; the mean improvement does not reach statistical significance. Appendix E.2 reports the complementary bootstrap analysis.

Transfer without adaptation. Figure 2a shows that EMOTRACK transfers more accurately than every trained baseline and retains performance close to the best prompted method. We freeze models trained and selected on DAIC-WOZ and evaluate them on E-DAIC without adaptation. EMOTRACK ranks second at 3.6977 ± 0.2665 , ahead of AIDA at 4.2538 and at least 1.21 MAE below the trained baselines, which range from 4.9047 to 5.3641. LMIQ (now) leads at 3.4722. Relative to in-domain evaluation, MAE rises by 1.23 for EMOTRACK, 1.43 for AIDA, and 0.44 for LMIQ. These results support the transfer value of explicit clinical evidence within a trainable predictor.

Longitudinal evaluation. Figure 2a shows that EMOTRACK maintains competitive accuracy across all three longitudinal sets, leading on LC8-GPT54-MINI and staying within 0.12 MAE of the best method on the other two. Each model is evaluated under the trajectory-level splits in Section 5.1; Appendix Table 8 reports all means and standard deviations. On LC8-GPT54-MINI, EMOTRACK reaches 3.8447 ± 0.1181 , improving on EnsemBERT (Qwen3-Emb) at 3.9184 ± 0.0753 by 1.9%. It obtains 2.6238 ± 0.0610 on LC8-QWEN35-35B, 0.0294 above Lau et al. (Qwen3-Emb), and 3.0887 ± 0.0740 on LC8-GPT56-LUNA, 0.1133 above Lau et al. The leading baseline changes across generators, while EMOTRACK is the only method in the top four on every evaluation set and in the top two on three of the five sets.

PHQ-8 item query readout. Table 2 shows that the PHQ-8 item queries improve total-score prediction on both primary benchmarks, supporting the value of separate item representations for total-score estimation. The control keeps the clinical tokens, client-turn tokens, encoder, and training recipe, and replaces the eight queries with a mean-pooled encoder head. This raises MAE from 2.4708 to 2.6889 on DAIC-WOZ and from 2.6238 to 2.7364 on LC8-QWEN35-35B. The 0.2181 increase on DAIC-WOZ is smaller than the 0.3526 gap to the strongest baseline. On LC8-GPT54-MINI the two heads differ by 0.0185 MAE. Gains are clearest on the real interview and larger longitudinal benchmarks.

Table 2: Effect of replacing the PHQ-8 item query readout with a mean-pooled encoder head on LC8-QWEN35-35B, DAIC-WOZ, and LC8-GPT54-MINI.

Prediction head	Overall MAE ↓		
	LC8-QWEN35-35B	DAIC-WOZ	LC8-GPT54-MINI
Mean-pooled encoder head	2.7364 ± 0.0411	2.6889 ± 0.1328	3.8262 ± 0.1046
PHQ-8 item query readout	2.6238 ± 0.0610	2.4708 ± 0.3892	3.8447 ± 0.1181

5.3 Ablation Study

Figure 2 and Tables 3 and 4 summarize the main ablations. Appendix F reports the remaining component comparisons and sweeps.

Input sources and speaker selection. Figures 2b and 2c show that combining clinical features with client-turn embeddings gives the lowest error on both primary benchmarks. The input-source ablation removes one branch at a time while keeping the encoder and readout; the speaker-selection ablation supplies client turns, counselor turns, or both. On DAIC-WOZ, removing clinical features raises MAE from 2.4708 to 4.7479, and removing embeddings raises it to 2.7049. On LC8-QWEN35-35B, embeddings alone reach 2.63, compared with 2.62 for both sources and 3.08 for clinical features alone. Clinical scores contribute most on the real interviews, while turn semantics contribute most on the longitudinal set. Adding counselor turns raises MAE to 2.85 and 2.71, and counselor-only input reaches 2.92 and 2.97. These controls support complementary evidence and client-focused inputs.

Predictor architecture. Table 3 shows that EMOTRACK outperforms standard regressors on identical upstream features, supporting the attention-based predictor. Ridge regression, an MLP,

and gradient-boosted trees receive the same 23 AIDA scores and Qwen3 embeddings for each seed. Gradient-boosted trees, the strongest of the three, reach 2.7881 on LC8-QWEN35-35B and 2.8070 on DAIC-WOZ, respectively 0.1643 and 0.3362 above EMOTRACK; the MLP and ridge regression are further behind. The gain therefore reflects more effective use of the same clinical–semantic inputs.

Table 3: Predictor ablation using identical AIDA features and Qwen3 embeddings.

Method	Overall MAE ↓	
	LC8-QWEN35-35B	DAIC-WOZ
Ridge regression	3.1709 ± 0.0027	3.1806 ± 0.0119
MLP regressor	2.9677 ± 0.0262	3.8103 ± 0.2072
Gradient-boosted trees	2.7881 ± 0.0092	2.8070 ± 0.1136
EMOTRACK	2.6238 ± 0.0610	2.4708 ± 0.3892

Auxiliary item-level supervision. Appendix Table 9 shows that the effect of auxiliary item supervision depends on the dataset. We vary λ_{sym} in Eq. (1) over $\{0, 0.1, 0.3, 0.5, 1.0\}$. On LC8-QWEN35-35B, the selected 0.5 reduces total-score MAE from 2.6481 to 2.6238 and item-level MAE from 0.5674 to 0.4251. On DAIC-WOZ, validation selects aggregate-only training, which reaches 2.4708 MAE, compared with 2.6937 at $\lambda_{\text{sym}} = 0.5$. On LC8-GPT54-MINI, aggregate-only training is also selected, illustrating the objective’s flexibility across supervision regimes.

Optional previous-session memory. Table 4 shows that the selected previous-session memory improves longitudinal accuracy on both evaluated sets, supporting compact use of historical context. We compare no memory, a summary path, and a combined summary and retrieval path using one prior session and the selected slot counts. Removing memory raises MAE from 2.6238 to 2.7097 on LC8-QWEN35-35B and from 3.8447 to 3.8631 on LC8-GPT54-MINI. Summary-only memory reaches 2.6036 and 3.9369, respectively. The preferred path depends on the setting; extending the same compressor to two prior sessions further lowers error on both sets (Appendix F.6). The benefit also depends on temporal continuity: memory improves estimates when the reference severity is unchanged, but increases MAE after changes of at least five PHQ-8 points in both longitudinal sets (Appendix G.3). The current gates therefore do not fully protect against outdated historical evidence.

Table 4: Optional-memory ablation on the two longitudinal sets. [†] marks the selected setting; bold marks the lowest test mean in each column.

Memory mechanism	Overall MAE ↓	
	LC8-QWEN35-35B	LC8-GPT54-MINI
No previous memory	2.7097 ± 0.0514	3.8631 ± 0.0779
Summary only	2.6036 ± 0.0432	3.9369 ± 0.1281
Summary + retrieval	2.6238 ± 0.0610 [†]	3.8447 ± 0.1181 [†]

6 Conclusion

We introduced EMOTRACK, which combines clinical features, frozen client-turn embeddings, and learned PHQ-8 item queries to estimate PHQ-8 total scores. Across five settings, it achieves the lowest normalized aggregate error, approximately 6.1% below the strongest baseline, and the lowest mean MAE on DAIC-WOZ, while remaining competitive under frozen transfer and longitudinal evaluation. An optional false-negative penalty yields the lowest missed screen-positive rate at a modest MAE cost. Ablations support the main design choices.

Limitations. Our evidence relies on synthetic longitudinal data and real single-session data, without clinical validation. Upstream models may not be the strongest available, and item-level predictions remain unsuitable for symptom-level use. Clinical use requires further calibration, triage thresholds, and human review. More detailed limitations are included in Appendix I.

References

- [1] Mohammad Amin Abbasi, Farnaz Sadat Mirnezami, Ali Neshati, and Hassan Naderi. HamRaz: A culture-based persian conversation dataset for person-centered therapy using llm agents. In *Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities*, pages 1–24, 2025.
- [2] Navneet Agarwal, Kirill Milintsevich, Lucie Metivier, Maud Rotharmel, Gaël Dias, and Sonia Dollfus. Analyzing symptom-based depression level estimation through the prism of psychiatric expertise. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 974–983, 2024.
- [3] Alexander Street. Counseling and psychotherapy transcripts: Volume i. Online database, n.d. URL <https://search.alexanderstreet.com/psyc>. Accessed September 5, 2026.
- [4] Suhas Bn, Dominik Mattioli, Andrew M Sherrill, Rosa I Arriaga, Christopher Wiese, and Saeed Abdullah. How real are synthetic therapy conversations? evaluating fidelity in prolonged exposure dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20986–20995, 2025. doi: 10.18653/v1/2025.findings-emnlp.1144.
- [5] Suhas BN, Andrew Sherrill, Rosa I. Arriaga, Christopher Wiese, and Saeed Abdullah. Thousand voices of trauma: A large-scale synthetic dataset for modeling prolonged exposure therapy conversations. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-2092. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5a29b161cf764f95e71488f07ecd64c0-Abstract-Datasets_and_Benchmarks_Track.html.
- [6] Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey, 2024. URL https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/DPQ_L.htm.
- [7] Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. Depression detection in clinical interviews with LLM-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8181–8194. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.452. URL <https://aclanthology.org/2024.naacl-long.452/>.
- [8] Sonia Curll, Kelly Mazzer, Sabina Albrecht, Skye Barbic, Amanda Fitzgerald, Kairi Kølves, Nic Telford, Nickolai Titov, and Debra Rickwood. Consensus derived client outcomes and clinician actions for youth online chat mental health services: A delphi study. *Frontiers in Digital Health*, 7:1671364, 2025. doi: 10.3389/fdgth.2025.1671364.
- [9] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 3123–3128, 2014.
- [10] Aylin Ece Gunal, Bowen Yi, John D. Piette, Rada Mihalcea, and Veronica Perez-Rosas. Examining Spanish counseling with MIDAS: a motivational interviewing dataset in Spanish. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 866–872, 2025. doi: 10.18653/v1/2025.naacl-short.73. URL <https://aclanthology.org/2025.naacl-short.73/>.
- [11] HPTN Socio-Behavioral and Structural Working Group. Depression measures working recommendations. Technical report, HIV Prevention Trials Network, 2022. URL https://www.hptn.org/sites/default/files/inline-files/HPTN%20SBSWG%20MHS%20Dep%20Meas%20Wkg%20Rec_V1.0_0.pdf. Version 1.0, 13 February 2022.
- [12] Hyunjong Kim, Suyeon Lee, Yeongjae Cho, Eunseo Ryu, Yohan Jo, Suran Seong, and Sungzoon Cho. Kmi: A dataset of korean motivational interviewing dialogues for psychotherapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10803–10828, 2025.
- [13] Subin Kim, Hoonrae Kim, Jihyun Lee, Yejin Jeon, and Gary Lee. Mirror: Multimodal cognitive reframing therapy for rolling with resistance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14851–14880, 2025.

- [14] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1–3):163–173, 2009. doi: 10.1016/j.jad.2008.06.026.
- [15] Xiaochong Lan, Zhiguang Han, Yiming Cheng, Li Sheng, Jie Feng, Chen Gao, and Yong Li. Depression detection on social media with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2155–2171, 2025.
- [16] Clinton Lau, Xiaodan Zhu, and Wai-Yip Chan. Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry*, 14:1160291, 2023.
- [17] Jae-Joong Lee, Jihoon Han, and Choong-Wan Woo. Interpretable depression assessment using a large language model. *PLOS Digital Health*, 5(2):e0001205, 2026.
- [18] Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, et al. CACTUS: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, 2024.
- [19] Jiayi Li, Zhaomin Wu, and Bingsheng He. LongCounsel-8: A benchmark suite for longitudinal depression tracking from multi-session counseling dialogues. *arXiv preprint arXiv:2609.03507*, 2026.
- [20] June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. Chatcounselor: A large language models for mental health support. *ArXiv*, abs/2309.15461, 2023.
- [21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
- [22] Mariko Makhmutova, Raghu Kainkaryam, Marta Ferreira, Jae Min, Martin Jaggi, and Ieuan Clay. Predicting changes in depression severity using the psyche-d (prediction of severity change-depression) model involving person-generated health data: longitudinal case-control observational study. *JMIR mHealth and uHealth*, 10(3):e34148, 2022.
- [23] Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 735–745. Association for Computing Machinery, 2022. doi: 10.1145/3488560.3498509. URL <https://doi.org/10.1145/3488560.3498509>.
- [24] Dhia Eddine Merzougui, Gaël Dias, Jeremie Pantin, and Fabrice Maurel. Evaluating large language models for depression symptom estimation. In *Artificial Intelligence in Medicine: 23rd International Conference, AIME 2025, Pavia, Italy, June 23-26, 2025, Proceedings, Part II*, pages 272–276, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-031-95840-3. doi: 10.1007/978-3-031-95841-0_51. URL https://doi.org/10.1007/978-3-031-95841-0_51.
- [25] Kirill Milintsevich, Kairit Sirts, and Gaël Dias. Towards automatic text-based estimation of depression through symptom prediction. *Brain informatics*, 10(1):4, 2023.
- [26] Subigya Nepal, Arvind Pillai, Weichen Wang, Tess Griffin, Amanda C Collins, Michael Heinz, Damien Lekkas, Shayan Mirjafari, Matthew Nemesure, George Price, et al. Moodcapture: Depression detection using in-the-wild smartphone images. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.
- [27] Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, Yougen Zhou, Yihan Li, Jingyuan Zhao, Qin Chen, et al. PsychEval: A multi-session and multi-therapy benchmark for high-realism AI psychological counselor. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 22192–22212, 2026. doi: 10.18653/v1/2026.findings-acl.1115.
- [28] Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. Kokorochat: A japanese psychological counseling dialogue dataset collected via role-playing by trained counselors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12424–12443, 2025. doi: 10.18653/v1/2025.acl-long.608. URL <https://aclanthology.org/2025.acl-long.608/>.
- [29] Huachuan Qiu and Zhenzhong Lan. Psydial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21624–21655, 2025.

- [30] Huachuan Qiu, Lizhi Ma, and Zhenzhong Lan. Psyguard: An automated system for suicide detection and risk assessment in psychological counseling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4607, 2024.
- [31] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL <https://qwen.ai/blog?id=qwen3.5>.
- [32] Federico Ravenda, Antonio Preti, Michele Poletti, Antonietta Mira, Fabio Crestani, and Andrea Raballo. Transforming social media text into predictive tools for depression through ai: A test-case study on the beck depression inventory-ii. *PLOS Digital Health*, 4(6):e0000848, 2025.
- [33] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop (AVEC)*, pages 3–12, 2019.
- [34] Gony Rosenman, Talma Hendler, and Lior Wolf. Llm questionnaire completion for automatic psychiatric assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 403–415, 2024.
- [35] 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*, 2024.
- [36] Aseem Srivastava, Tharun Suresh, S. P. Lord, Md. Shad Akhtar, and Tanmoy Chakraborty. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3920–3930, 2022. doi: 10.1145/3534678.3539187.
- [37] Xiaoyi Wang, Jiwei Zhang, Guangtao Zhang, and Honglei Guo. Feel the difference? a comparative analysis of emotional arcs in real and LLM-generated CBT sessions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19999–20017. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1089. URL <https://aclanthology.org/2025.findings-emnlp.1089/>.
- [38] Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.clpsych-1.8. URL <https://aclanthology.org/2024.clpsych-1.8/>.
- [39] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-MI: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9746035. URL <https://doi.org/10.1109/ICASSP43922.2022.9746035>.
- [40] Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115, 2025.
- [41] Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost B. Wagenaar, George Demiris, and Li Shen. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V2*, pages 5367–5378, 2025. doi: 10.1145/3711896.3737393.
- [42] Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. Cpsychoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, 2024.
- [43] Linhai Zhang, Ziyang Gao, Deyu Zhou, and Yulan He. Explainable depression detection in clinical interviews with personalized retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9927–9944, 2025.

- [44] Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 146–158, 2024.
- [45] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. doi: 10.48550/arXiv.2506.05176. URL <https://arxiv.org/abs/2506.05176>.
- [46] Rou Zhong, Nanxi Li, Le Xiao, Lei Feng, Yuan Feng, Gang Wang, and Xuequan Zhu. Contribution of longitudinal mobile health measures in the dynamic track of patients with major depressive disorder: Multiple centers, prospective cohort study using functional data analysis and machine learning. *JMIR mHealth and uHealth*, 14:e81397, 2026.

Appendix

Table of Contents

A Additional Related Work	14
A.1 Transcript-Based Depression Severity Prediction	14
A.2 Counseling and Motivational Interviewing Datasets	15
B Algorithmic Details of EMOTRACK	15
C Benchmark Protocols	15
C.1 LongCounsel-8	15
C.2 Frozen Transfer from DAIC-WOZ to E-DAIC	16
D Additional Experiment Settings	16
E Complete Method Comparison	18
E.1 Results by Set	18
E.2 Participant-Level Robustness on DAIC-WOZ	19
E.3 EnsemBERT Hyperparameter Selection	19
F Ablation Studies	19
F.1 Auxiliary Item-Level Supervision	19
F.2 Encoder–Decoder Depth	20
F.3 History Dropout	20
F.4 Maximum Context Sentences	20
F.5 Memory Slots	21
F.6 History Scope	21
F.7 Clinical-Feature Extractor	22
G Further Analyses	22
G.1 False-Negative Penalty	22
G.2 Per-Item Error and Directional Bias	22
G.3 Session-to-Session Transitions	23
G.4 Performance by Trajectory Type	23
G.5 Runtime and Inference Cost	23
H Asset Licenses and Intended Use	24
I Limitations	25

A Additional Related Work

A.1 Transcript-Based Depression Severity Prediction

Table 5 provides a detailed comparison of text-based depression-evaluation methods that produce symptom-severity scores.

Table 5: Approaches to depression evaluation.

Study	Modality	Data	Clinical Criteria	Output
<i>Multimodal methods</i>				
MoodCapture [26]	Image	Smartphone images	PHQ-8, DSM-5	Score
LLM+Landmarks [44]	Speech + Text	DAIC-WOZ	PHQ-8	Binary
Sadeghi et al. [35]	Video + Speech + Text	E-DAIC	PHQ-8	Score
SEGA [7]	Video + Speech + Text	DAIC-WOZ	PHQ-8	Binary
<i>Text-based methods with classification output</i>				
DORIS [15]	Text	Weibo & Twitter posts	DSM-5	Binary
Wang et al. [38]	Text	Social-media posts	BDI	Score
RED [43]	Text	DAIC-WOZ	PHQ-8	Binary
PsyGUARD [30]	Text	Counseling utterances	C-SSRS	Level
<i>Text-based methods with score output</i>				
Lau et al. [16]	Text	DAIC-WOZ	PHQ-8	Score
PsychiatricPrism [2]	Text	DAIC-WOZ	PHQ-8	Score
LMIQ [34]	Text	DAIC-WOZ	PHQ-8	Score
AIDA [17]	Text	DAIC-WOZ	PHQ-8	Score
Milintsevich et al. [25]	Text	DAIC-WOZ	PHQ-8	Score
EnsemBERT [32]	Text	Reddit posts	BDI-II	Score
EMOTRACK (ours)	Text	DAIC-WOZ, E-DAIC & LongCounsel-8	PHQ-8	Score

A.2 Counseling and Motivational Interviewing Datasets

To complement Section 2, Table 6 surveys counseling and motivational-interviewing resources by provenance, language, multi-session structure, and PHQ supervision. Real corpora are drawn from demonstrations, role-play, reconstructed sessions, or licensed transcripts [3, 10, 20, 23, 28, 29, 36, 37, 39]; synthetic corpora use varied therapeutic frameworks [1, 4, 5, 12, 13, 18, 27, 40, 42]. DAIC-WOZ [9] supplies real single-session PHQ labels, and LongCounsel-8 [19] supplies the multi-session evaluation used here.

B Algorithmic Details of EMOTRACK

Algorithm 1 summarizes the forward computation of EMOTRACK. Section 4 defines the clinical-feature representation module, dialogue representation module, Transformer backbone, and previous-session memory module. The algorithm uses these components to express the complete forward pass.

C Benchmark Protocols

This appendix records the properties and protocols of the LongCounsel-8 benchmark and of the E-DAIC transfer set that are needed to interpret the experiments. The benchmark paper [19] documents the generation prompts and validity audits.

C.1 LongCounsel-8

LongCounsel-8 [19] is a synthetic benchmark for longitudinal PHQ-8 assessment from counseling dialogue. Each trajectory spans five counseling sessions, and each session contains 20 counselor and 20 client turns. Client profiles and session structure derive from PsychEval cases [27], five-visit severity trajectories follow the PSYCHE-D study [22], symptom co-occurrence follows NHANES DPQ responses [6], and dialogue form is calibrated against RealCBT [37]. Each session is generated from a latent PHQ-8 state comprising eight item values and is paired with a simulated self-report label. The latent values condition dialogue generation but are not exposed as transcript input. Following the benchmark protocol, models are trained with the simulated self-report labels and evaluated against the

Table 6: Counseling and motivational interviewing dialogue resources.

Work	Type	Language	Multi-session	PHQ-supervised
<i>Real counseling / motivational interviewing resources</i>				
KokoroChat [28]	Real	JA	✗	✗
HOPE [23]	Real	EN	✗	✗
MEMO [36]	Real	EN	✗	✗
Psych8K [20]	Real	EN	✗	✗
PsyDial-D0_m [29]	Real	ZH	✓	✗
Counseling and Psychotherapy Transcripts [3]	Real	EN	✓	✗
RealCBT [37]	Real	EN	✗	✗
AnnoMI [39]	Real	EN	✗	✗
MIDAS [10]	Real	ES	✗	✗
MentalChat16K [41]	Mixed	EN	✗	✗
DAIC-WOZ [9]	Real	EN	✗	✓
<i>Synthetic counseling / motivational interviewing resources</i>				
PsyDT / PsyDTCorpus [40]	Synthetic	ZH	✗	✗
PsyDial [29]	Synthetic	ZH	✗	✗
CACTUS [18]	Synthetic	EN	✗	✗
CPsyCoun / Memo2Demo [42]	Synthetic	ZH	✗	✗
HamRaz [1]	Synthetic	FA	✗	✗
MIRROR [13]	Synthetic	EN	✗	✗
KMI [12]	Synthetic	KO	✗	✗
Thousand Voices of Trauma [5]	Synthetic	EN	✗	✗
PsychEval [27]	Synthetic	ZH	✓	✗
LongCounsel-8 [19]	Synthetic	EN	✓	✓

latent state, requiring them to infer the reference severity from the dialogue. Because the benchmark is synthetic, its results are controlled benchmark evidence rather than clinical validation (Appendix I).

We use three releases. LC8-QWEN35-35B is the main set: 3,690 trajectories (18,450 sessions) generated with Qwen3.5-35B-A3B [31], with a trajectory-level split of 2,583 training, 369 validation, and 738 test trajectories. LC8-GPT54-MINI and LC8-GPT56-LUNA are two 369-trajectory sets (1,845 sessions each) generated with GPT-5.4-mini and GPT-5.6 Luna, with one profile-trend trajectory per source profile and a 70/10/20 trajectory-level split. The two smaller sets share scale and protocol and differ only in generator, which isolates the generator effect; LC8-GPT54-MINI differs from LC8-QWEN35-35B in both generator and scale and serves as a joint stress test on both. Assigning all five sessions of a trajectory to one partition prevents cross-session leakage. We use the released transcripts, labels, and split manifests.

C.2 Frozen Transfer from DAIC-WOZ to E-DAIC

E-DAIC [33] extends the DAIC-WOZ interview protocol to further participants and collection waves. For the transfer set, every model is trained and selected on the official DAIC-WOZ splits and then applied to E-DAIC without adaptation; E-DAIC labels are used only for scoring. The setting therefore measures out-of-the-box generalization to a new collection under the same protocol.

D Additional Experiment Settings

Hyperparameters. Table 7 summarizes the shared architecture and default EMOTRACK hyperparameters, with dataset-specific settings described here. On LC8-GPT54-MINI, validation selects $\lambda_{\text{sym}}=0$, $S=16$, and $p_{\text{hist}}=0.2$; on LC8-QWEN35-35B the corresponding values are 0.5, 4, and 0.1. On DAIC-WOZ, validation selects three encoder layers, four decoder layers, and $\lambda_{\text{sym}}=0$; frozen

Algorithm 1 EMOTRACK forward computation

Require: Current transcript $\mathcal{T}^{(t)}$; optional previous transcript $\mathcal{T}^{(t-1)}$; learned PHQ-8 item queries $\mathbf{Q} \in \mathbb{R}^{J \times d}$, where $J = 8$

Ensure: Predicted total PHQ-8 score $\hat{Y} \in [0, 24]$

- 1: **// Current-session representation**
- 2: $\mathbf{a}^{(t)} \leftarrow \phi_{\text{LLM}}(\mathcal{T}^{(t)})$
- 3: $\mathbf{H}^c \leftarrow \Phi_{\text{clin}}(\mathbf{a}^{(t)})$
- 4: $\mathbf{S}^{(t)} \leftarrow \Phi_{\text{dlg}}(E(\mathcal{T}^{(t)}))$
- 5: $\mathbf{X}^{(t)} \leftarrow [\mathbf{H}^c; \mathbf{S}^{(t)}]$
- 6: **// Current-session PHQ-8 item encoding**
- 7: $\mathbf{D}^{(t)} \leftarrow \mathcal{B}_{\theta}(\mathbf{X}^{(t)}, \mathbf{Q})$
- 8: **// Optional previous-session memory refinement**
- 9: **if** $\mathcal{T}^{(t-1)}$ is available and not dropped by history dropout **then**
- 10: $\mathbf{V}^{(t-1)} \leftarrow \Phi_{\text{hist}}(E(\mathcal{T}^{(t-1)}))$
- 11: $\mathbf{D}^{(t)} \leftarrow \mathcal{R}_{\theta}(\mathbf{D}^{(t)}, \mathbf{V}^{(t-1)})$
- 12: **end if**
- 13: **// Total-score readout**
- 14: $\hat{y}_j \leftarrow 3 \cdot \sigma(\mathbf{w}_j^{\top} \mathbf{d}_j^{(t)} + b_j), \quad j = 1, \dots, J$
- 15: $\hat{Y} \leftarrow \sum_{j=1}^J \hat{y}_j$
- 16: **return** $\hat{Y}$

E-DAIC transfer retains this configuration. Previous-session memory is omitted in both settings. The model has $\sim 1.02\text{M}$ trainable parameters.

Table 7: Default hyperparameters for EMOTRACK; dataset-specific settings are described in the text.

Component	Value	Component	Value
Model dimension d	64	AIDA features F	23
Attention heads h	4	Dropout	0.1
Feed-forward dimension	256	History dropout p_{hist}	0.1
Encoder layers L_{enc}	2	Learning rate	10^{-3}
Decoder layers L_{dec}	4	Weight decay	10^{-2}
Memory slots S	4	Batch size	32
Score MLP hidden dim	32	Warmup ratio	0.05
DAIC max utterances N_{max}	80	Symptom-loss weight λ_{sym}	0.5
PHQ-8 symptoms J	8	Penalty weight λ_{FN}	0 (2.5 for screening)

Environment. All experiments were performed on a server running Ubuntu 22.04 LTS, equipped with an AMD EPYC 9654 96-core processor (192 threads), 1.1 TiB of system memory, and 4 NVIDIA H100 PCIe GPUs with 80 GiB HBM each (320 GiB aggregate GPU memory).

Baseline implementations. We implement baselines from released algorithms or author code when available, and use prior DAIC-WOZ results only as reproducibility anchors because published subsets, preprocessing, checkpoint selection, and scoring rules can differ. **AIDA [17] and LMIQ [34] align with prior reports:** AIDA obtains 2.8234 ± 0.2099 MAE in our five-seed sweep, compared with 2.91 in Lee et al. [17]; LMIQ obtains 3.0285 ± 0.0460 , compared with 3.46 in Rosenman et al. [34].

Lau et al. [16] and Milintsevich et al. [25] are more implementation-sensitive. Lau et al. [16] report 3.80 test MAE on DAIC-WOZ, whereas our author-code run with the released dual-encoder configuration and all-mpnet-base-v2 obtains 4.4367 ± 0.5744 . Seed 4 reaches 3.9680, while seed 0 selects an epoch-0 checkpoint with near-constant predictions and yields 5.4202 MAE. The Qwen3-Embedding variant shows a similar pattern: one seed reaches 3.6330, but the five-seed mean remains 4.0927 ± 0.2795 . Milintsevich et al. [25] report 3.78 ± 0.13 test MAE, whereas our five-seed sweep obtains 4.4370 ± 0.1197 (best seed 4.2957) and our direct author-repository audit obtains 4.5164. Because all five seeds remain above the reported score, seed choice alone does not explain the discrepancy. The audit instead indicates sensitivity to preprocessing and checkpoint selection,

including exclusion of one training participant with missing item labels and two invalid development transcripts, development-loss checkpoint selection, and aggregate-only PHQ-8 test evaluation.

EnsemBERT [32] is an architecture-level adaptation. Ravenda et al. [32] evaluate BDI-II prediction from Reddit/eRisk post histories rather than PHQ-8 prediction on DAIC-WOZ. Under its published hyperparameters, two of the five all-mpnet-base-v2 seeds on the small DAIC-WOZ training split produce all-zero predictions with MAE 6.9787, giving 5.6681 ± 1.2407 overall. We therefore re-select its optimization and capacity hyperparameters on the official DAIC-WOZ validation split and report the selected setting on every evaluation set (Appendix E.3).

E Complete Method Comparison

Table 8 shows that EMOTRACK combines strong performance across all five evaluation sets, while Figure 2a summarizes its leading normalized aggregate error. Section 5.1 and Appendix C describe the five sets.

Table 8: Total-score MAE of every method on the five evaluation sets (five-seed mean $\pm$ SD; bold marks the lowest mean in each column). The last column applies the DAIC-WOZ-trained models to E-DAIC without adaptation.

Method	Overall MAE $\downarrow$				
	LC8-QWEN35-35B	LC8-GPT54-MINI	LC8-GPT56-LUNA	DAIC-WOZ	DAIC-WOZ $\rightarrow$ E-DAIC
AIDA	3.2210	4.0689	3.3699	2.8234	4.2538
(now)	± 0.0068	± 0.0133	± 0.0117	± 0.2099	± 0.2697
AIDA	3.2517	4.0948	3.4290	—	—
(all)	± 0.0051	± 0.0243	± 0.0157	—	—
LMIQ	2.9792	4.0095	3.3032	3.0285	3.4722
(now)	± 0.0117	± 0.0566	± 0.0608	± 0.0460	± 0.0729
LMIQ	2.9585	3.9908	3.0585	—	—
(all)	± 0.0185	± 0.0513	± 0.0541	—	—
Milintsevich	2.6953	4.1530	3.3494	4.4370	5.0497
et al.	± 0.0211	± 0.0602	± 0.0271	± 0.1197	± 0.1233
EnsemBERT	2.8618	4.0038	3.3400	4.1149	4.9047
	± 0.0709	± 0.1396	± 0.1161	± 0.3339	± 0.3487
EnsemBERT	2.7492	3.9184	3.1535	4.8255	5.2465
(Qwen3-Emb)	± 0.0343	± 0.0753	± 0.0382	± 0.7277	± 0.5317
Lau et al.	2.6204	4.0689	2.9754	4.4367	5.1880
	± 0.0194	± 0.0865	± 0.1045	± 0.5744	± 0.3158
Lau et al.	2.5944	4.0640	3.0160	4.0927	5.3641
(Qwen3-Emb)	± 0.0232	± 0.1034	± 0.0448	± 0.2795	± 0.2134
EMOTRACK	2.6238	3.8447	3.0887	2.4708	3.6977
	± 0.0610	± 0.1181	± 0.0740	± 0.3892	± 0.2665

E.1 Results by Set

Table 8 shows that EMOTRACK has the lowest mean MAE on DAIC-WOZ and LC8-GPT54-MINI and stays within 0.2255 MAE of the leading method on the other three sets. On real DAIC-WOZ, EMOTRACK reaches 2.4708 ± 0.3892 against 2.8234 ± 0.2099 for AIDA, the strongest baseline, a 12.5% reduction. On LC8-GPT54-MINI it reaches 3.8447 ± 0.1181 against 3.9184 ± 0.0753 for EnsemBERT (Qwen3-Emb), a margin of 0.0737 MAE (1.9%). It ranks third on LC8-QWEN35-35B at 2.6238 ± 0.0610 , 0.0294 behind Lau et al. (Qwen3-Emb) at 2.5944 ± 0.0232 and 0.0034 behind the original Lau encoder; fourth on LC8-GPT56-LUNA at 3.0887 ± 0.0740 , 0.1133 behind Lau et al. at 2.9754 ± 0.1045 ; and second on E-DAIC at 3.6977 ± 0.2665 , 0.2255 behind LMIQ (now) at 3.4722 ± 0.0729 and ahead of every trained baseline by at least 1.21 MAE.

Two patterns follow. First, the ordering among methods depends on the set: the trained transcript encoders of Lau et al. lead on the Qwen- and Luna-generated corpora, while the prompted clinical methods AIDA and LMIQ hold up best under distribution shift, which places LMIQ (now) first under

frozen E-DAIC transfer. EMOTRACK draws on both evidence sources and is the only method that ranks first or second on three of the five sets and within the top four on all of them. Second, the two 369-trajectory sets order the methods differently while sharing scale and protocol; generator identity alone therefore moves the ranking, and reporting all three synthetic sets gives a fuller picture than any one of them.

E.2 Participant-Level Robustness on DAIC-WOZ

The participant-level analysis extends the comparison in Table 8: EMOTRACK improves on AIDA for most DAIC-WOZ test participants, while the bootstrap interval leaves the population-level improvement uncertain. We average each participant’s prediction over the five seeds before scoring, compare absolute errors participant by participant against AIDA, and bootstrap over participants for a 95% confidence interval. EMOTRACK reduces MAE by 0.3324 and has the lower absolute error for 28 of 47 participants (59.6%); the interval for the reduction is $[-0.1581, 0.8154]$. The reduction of 0.3324 is smaller than the 0.3526 difference of five-seed means in Table 8 because averaging predictions within each participant removes seed variance that the mean of means retains; the two quantities answer different questions. The positive average difference and the majority of improved participants support the observed benchmark gain; the interval spanning zero is consistent with the paired permutation result of $p = 0.1901$.

E.3 EnsemBERT Hyperparameter Selection

Re-selecting EnsemBERT’s hyperparameters on the DAIC-WOZ validation split lowers its DAIC-WOZ test MAE from 5.6681 ± 1.2407 to 4.1149 ± 0.3339 but does not change its ranking relative to EMOTRACK (Table 8). EnsemBERT was developed for BDI-II estimation from Reddit/eRisk post histories, and its published hyperparameters (learning rate 10^{-3} , no weight decay, 20 retrieved posts, 128 dense units, one attention head, batch size 32) give constant predictions on several DAIC-WOZ seeds. We therefore re-selected these six hyperparameters over three values each (729 configurations, five seeds each) using only the official DAIC-WOZ training and validation splits; the architecture, the frozen all-mpnet-base-v2 representation model, the retrieval model, and the loss weights were held fixed, and the test split was opened once after the winning configuration (learning rate 10^{-3} , weight decay 10^{-4} , 30 posts, 256 units, two heads, batch size 64) was frozen. Selection removes the constant-output seeds and cuts the cross-seed spread from 1.2407 to 0.3339. The same configuration is then applied unchanged to the other four sets and to the Qwen3-Embedding variant, which makes those values a conservative estimate of per-dataset selection. The retuned EnsemBERT remains fifth of eight methods on DAIC-WOZ, and although its Qwen3-Embedding variant becomes the strongest baseline on LC8-GPT54-MINI (3.9184), EMOTRACK still has the lower mean (3.8447).

F Ablation Studies

Section 5.3 reports the input-source, speaker-selection, predictor, item-supervision, and memory studies. This appendix gives the remaining ablations. [†] marks the selected setting for each dataset. Bold marks the lowest test mean within each axis and dataset; these descriptive minima summarize sensitivity to each component.

F.1 Auxiliary Item-Level Supervision

Table 9 gives the full sweep behind the item-supervision result of Section 5.3: item-level MAE on LC8-QWEN35-35B falls monotonically with the weight, from 0.5674 at $\lambda_{\text{sym}}=0$ to 0.4178 at 1.0, while aggregate MAE is lowest at the selected 0.5 on LC8-QWEN35-35B and 0 on DAIC-WOZ. We vary λ_{sym} in Eq. (1) over $\{0, 0.1, 0.3, 0.5, 1.0\}$ with all other settings fixed. On LC8-QWEN35-35B the aggregate means stay within 0.03 of each other (2.6238 to 2.6519), and on DAIC-WOZ aggregate-only training reaches 2.4708 against 2.6937 at 0.5 and 2.6477 at 1.0. On LC8-GPT54-MINI, aggregate-only training is selected (3.8447) while the post-hoc test minimum is 3.7454 at 0.3. Auxiliary item supervision therefore improves item-level accuracy on LC8-QWEN35-35B, while its benefit for total-score prediction depends on the dataset.

Table 9: Auxiliary item-level-loss ablation. [†] marks the selected setting; bold marks the lowest test mean in each column.

Objective	Aggregate MAE ↓			Item MAE ↓
	LC8-QWEN35-35B	DAIC-WOZ	LC8-GPT54-MINI	LC8-QWEN35-35B
$\lambda_{\text{sym}} = 0$ (aggregate only)	2.6481±0.0370	2.4708±0.3892 [†]	3.8447±0.1181 [†]	0.5674±0.0041
$\lambda_{\text{sym}} = 0.1$	2.6519±0.0555	2.5600±0.2717	3.8565±0.1603	0.4847±0.0206
$\lambda_{\text{sym}} = 0.3$	2.6372±0.0492	2.5785±0.1866	3.7454±0.1019	0.4310±0.0125
$\lambda_{\text{sym}} = 0.5$	2.6238±0.0610 [†]	2.6937±0.2892	3.7509±0.0578	0.4251±0.0124
$\lambda_{\text{sym}} = 1.0$	2.6271±0.0442	2.6477±0.2494	3.8870±0.1206	0.4178±0.0131

F.2 Encoder–Decoder Depth

Table 10 shows that the selected depths, 2–4 on LC8-QWEN35-35B and LC8-GPT54-MINI and 3–4 on DAIC-WOZ, are competitive with shallower and deeper variants, with every test mean within 0.16 MAE of the selected setting. We vary the numbers of encoder and decoder layers over five combinations with all other hyperparameters fixed. On LC8-QWEN35-35B, the selected setting obtains 2.6238±0.0610, while the five test means range only from 2.6071 (2–6) to 2.6576 (2–2). On DAIC-WOZ, the selected 3–4 setting has the lowest test mean (2.4708±0.3892), compared with 2.5125 for 1–4 and 2.6241 for 2–4. On LC8-GPT54-MINI, the selected setting reaches 3.8447±0.1181, only 0.0444 above the descriptive minimum at 1–4. The selected moderate-depth backbone supports accurate prediction across the three evaluation sets.

Table 10: Encoder–decoder depth on LC8-QWEN35-35B, DAIC-WOZ, and LC8-GPT54-MINI.

Layers		Overall MAE ↓		
Encoder	Decoder	LC8-QWEN35-35B	DAIC-WOZ	LC8-GPT54-MINI
1	4	2.6314 ± 0.0371	2.5125 ± 0.2239	3.8003 ± 0.0916
2	2	2.6576 ± 0.0674	2.6274 ± 0.3043	3.8790 ± 0.2592
2	4	2.6238 ± 0.0610 [†]	2.6241 ± 0.2208	3.8447 ± 0.1181 [†]
2	6	2.6071 ± 0.0362	2.6053 ± 0.2492	3.8648 ± 0.0460
3	4	2.6163 ± 0.0304	2.4708 ± 0.3892 [†]	3.8615 ± 0.1363

F.3 History Dropout

Table 11 supports light history dropout: the selected rates outperform the strongest tested rate on both longitudinal sets. We vary p_{hist} over $\{0, 0.1, 0.2, 0.3\}$ with the memory module otherwise unchanged. On LC8-QWEN35-35B, the selected 0.1 is also the lowest-test-MAE setting at 2.6238±0.0610; removing dropout or raising it to 0.2 and 0.3 gives 2.6824, 2.6434, and 2.6744. On LC8-GPT54-MINI, the selected 0.2 reaches 3.8447±0.1181, while the descriptive minimum occurs without dropout (3.7898) and 0.3 performs worst (3.9462). These results favor retaining useful prior context with a light perturbation rate.

Table 11: Effect of history dropout on LC8-QWEN35-35B and LC8-GPT54-MINI.

History dropout	Overall MAE ↓	
	LC8-QWEN35-35B	LC8-GPT54-MINI
0.0	2.6824 ± 0.0467	3.7898 ± 0.1085
0.1	2.6238 ± 0.0610 [†]	3.8057 ± 0.0622
0.2	2.6434 ± 0.0735	3.8447 ± 0.1181 [†]
0.3	2.6744 ± 0.0650	3.9462 ± 0.1312

F.4 Maximum Context Sentences

On DAIC-WOZ, keeping the first 80 participant utterances is both the selected and the lowest-test-MAE cap (Table 12). We vary N_{max} over $\{20, 40, 80, 384\}$ on DAIC-WOZ, whose interviews

frequently exceed 80 utterances; the longitudinal sets are excluded because every session contains only 20 client turns, all of which are retained under every tested cap. Test MAE falls from 2.7568 ± 0.2008 at 20 to 2.7236 ± 0.3455 at 40 and 2.4708 ± 0.3892 at 80, and rises again to 2.5903 ± 0.2637 at 384. This comparison favors the first 80 client utterances for the evaluated model.

Table 12: Effect of the maximum number of retained participant utterances on DAIC-WOZ.

Maximum participant utterances	DAIC-WOZ Overall MAE ↓
20	2.7568 ± 0.2008
40	2.7236 ± 0.3455
80	2.4708 ± 0.3892 †
384	2.5903 ± 0.2637

F.5 Memory Slots

Table 13 shows that compact memory is sufficient: four slots achieve the lowest test error on LC8-QWEN35-35B, while the four tested means on LC8-GPT54-MINI span only 0.0743. We vary the number of slots S over $\{4, 8, 12, 16\}$ with the rest of the memory module fixed. On LC8-QWEN35-35B, the selected four-slot setting also has the lowest test MAE (2.6238), while 8, 12, and 16 slots give 2.6607, 2.6617, and 2.6975. LC8-GPT54-MINI selects 16 slots (3.8447), although its descriptive minimum occurs at four slots (3.8058). The small spread supports a compact memory representation with capacity selected for each setting.

Table 13: Effect of memory capacity on LC8-QWEN35-35B and LC8-GPT54-MINI.

Memory slots	Overall MAE ↓	
	LC8-QWEN35-35B	LC8-GPT54-MINI
4	2.6238 ± 0.0610 †	3.8058 ± 0.0926
8	2.6607 ± 0.0946	3.8801 ± 0.1487
12	2.6617 ± 0.0416	3.8317 ± 0.0925
16	2.6975 ± 0.0417	3.8447 ± 0.1181 †

F.6 History Scope

Two prior sessions give the lowest MAE on both longitudinal sets, and full history stays competitive, which shows that the fixed-slot compressor is robust to the history scope (Table 14). Section 5.3 compares memory *paths* at a fixed one-session scope; here we vary the *scope*, extending the fixed-slot compressor to two prior sessions and to the full available history under identical splits and hyperparameters, with the selected prior client-turn embeddings entering the unchanged compressor and past labels withheld. Two prior sessions reach 2.5880 on LC8-QWEN35-35B and 3.7559 on LC8-GPT54-MINI, improving on the deployed one-session setting by 0.0358 and 0.0888; full history stays close at 2.6111 and 3.7870, only 0.0231 and 0.0311 behind. Longer histories therefore fit the compressor as well as short ones, with two sessions the best-performing choice among those tested.

Table 14: Effect of history scope on the two multi-session sets, with the previous-session compressor unchanged. Bold marks the lowest mean in each column.

History scope	Overall MAE ↓	
	LC8-QWEN35-35B	LC8-GPT54-MINI
No previous memory	2.7097 ± 0.0514	3.8631 ± 0.0779
Previous 1 session	2.6238 ± 0.0610	3.8447 ± 0.1181
Previous 2 sessions	2.5880 ± 0.0808	3.7559 ± 0.0977
All history	2.6111 ± 0.0742	3.7870 ± 0.1314

F.7 Clinical-Feature Extractor

EMOTRACK has lower mean MAE than the matched AIDA regressor under two clinical-feature extractors in a separate reference configuration. To examine dependence on the extractor noted in Appendix I, we compare Qwen and GPT-5.6 Luna extraction on the same DAIC-WOZ transcripts over five seeds, using two encoder layers, four decoder layers, $\lambda_{\text{sym}}=0.5$, and a cap of 80 client utterances. Under Luna extraction, EMOTRACK obtains 2.6548 ± 0.1057 against 2.8091 ± 0.1003 for AIDA and 3.1223 ± 0.0508 for LMIQ; under Qwen extraction the values are 2.5831 ± 0.2332 , 2.8234 ± 0.2099 , and 3.0285 ± 0.0460 . EMOTRACK therefore leads under either extractor, retaining a 0.1543 MAE advantage over AIDA with Luna features; its own error rises by 0.0717 between these reference runs. The Qwen and Luna runs differ in the deterministic-training setting, so this comparison is descriptive and does not fully isolate extractor effects.

G Further Analyses

This appendix collects analyses of EMOTRACK’s screening behavior, item-level errors, longitudinal behavior, and inference cost.

G.1 False-Negative Penalty

Table 15 shows that the optional penalty gives EMOTRACK the **lowest false-negative rate among the evaluated methods**, supporting explicit control over missed screen-positive sessions. We evaluate on LC8-GPT54-MINI using MAE, FNR, and FPR. We train EMOTRACK with $\lambda_{\text{FN}} \in \{0, 0.5, 1, 2, 2.5, 3, 4, 5\}$ in Eq. (2) and select 2.5 by validation FNR. At the selected weight, FNR falls from 0.7080 ± 0.0694 to 0.4220 ± 0.0853 , compared with 0.5880 for the strongest baseline, Lau et al. MAE rises from 3.8447 to 4.0370 ± 0.0691 and FPR from 0.1111 to 0.2667. The penalty therefore reduces missed positives at the cost of some score accuracy and additional false positives.

Table 15: Screening comparison on LC8-GPT54-MINI at the PHQ-8 threshold of 10; bold marks the best value in each column.

Method	FNR ↓	FPR ↓	MAE ↓
AIDA (now)	0.7500 ± 0.0000	0.0778 ± 0.0037	4.0689 ± 0.0133
AIDA (all)	0.7300 ± 0.0187	0.0859 ± 0.0088	4.0948 ± 0.0243
LMIQ (now)	0.7040 ± 0.0167	0.0963 ± 0.0159	4.0095 ± 0.0566
LMIQ (all)	0.7220 ± 0.0349	0.0793 ± 0.0165	3.9908 ± 0.0513
Milintsevich et al.	0.6460 ± 0.0503	0.1481 ± 0.0183	4.1530 ± 0.0602
EnsemBERT	0.7560 ± 0.1704	0.0956 ± 0.0899	4.0038 ± 0.1396
EnsemBERT (Qwen3-Emb)	0.6620 ± 0.1182	0.1370 ± 0.0522	3.9184 ± 0.0753
Lau et al.	0.5880 ± 0.0638	0.1681 ± 0.0330	4.0689 ± 0.0865
Lau et al. (Qwen3-Emb)	0.6060 ± 0.0336	0.1756 ± 0.0279	4.0640 ± 0.1034
EMOTRACK, $\lambda_{\text{FN}}=0$	0.7080 ± 0.0694	0.1111 ± 0.0372	3.8447 ± 0.1181
EMOTRACK, $\lambda_{\text{FN}}=2.5$	0.4220 ± 0.0853	0.2667 ± 0.0515	4.0370 ± 0.0691

The weight sweep shows that a moderate penalty captures the screening benefit. FNR already falls to 0.5680 at $\lambda_{\text{FN}}=0.5$ before reaching 0.4220 at 2.5. Beyond 2.5, FNR stays between 0.4360 and 0.4840 while MAE reaches as much as 4.3872; larger weights therefore increase error without reducing missed positives.

G.2 Per-Item Error and Directional Bias

We examine systematic under- and over-prediction through mean signed error. Every PHQ-8 item stays below one point of MAE on the DAIC-WOZ development split, and the signed biases are directionally consistent: the somatic items are underestimated and psychomotor change is overestimated (Table 16). We score EMOTRACK’s intermediate item values against the released item labels of the 35-participant development split; the public test split lacks item-level labels. Item MAE ranges from 0.4707 for anhedonia to 0.7539 for sleep. The bias is negative for sleep (-0.3428), fatigue (-0.4586), and appetite (-0.2293) and positive for psychomotor change ($+0.4554$) and anhedonia ($+0.2012$). These biases partly offset in the sum, which leaves the total-score results intact and

matches the role of the PHQ-8 item query readout as an architectural prior on the total score; they also mark where item-wise calibration would be needed for symptom-level use, with the somatic items as the natural first target.

Table 16: Per-symptom MAE and signed bias (prediction minus label) of EMOTRACK on the 35-participant DAIC-WOZ development split, the only partition with released item-level labels.

PHQ-8 symptom	Item MAE ↓	Signed bias
Anhedonia	0.4707	+0.2012
Depressed mood	0.5191	−0.0871
Sleep	0.7539	−0.3428
Fatigue	0.6990	−0.4586
Appetite	0.6931	−0.2293
Self-worth	0.6618	−0.1438
Concentration	0.6798	+0.0289
Psychomotor	0.6226	+0.4554

G.3 Session-to-Session Transitions

Previous-session memory improves estimation when reference severity is unchanged, consistent with a benefit from temporal continuity. We stratify test sessions by the change in the latent PHQ-8 total from the previous session and compare EMOTRACK with and without one-session memory. When the prior state is unchanged, memory reduces MAE from 2.4066 to 2.1852 on LC8-QWEN35-35B and from 4.0838 to 3.8426 on LC8-GPT54-MINI. After a change of at least 5 PHQ points the ordering reverses: without memory the model reaches 3.5480 on LC8-QWEN35-35B and 4.8127 on LC8-GPT54-MINI, against 3.6258 and 4.9745 with memory. These comparisons show that historical context can help when severity persists but can harm prediction after an abrupt transition. The current gates do not fully suppress this effect, motivating future mechanisms that explicitly account for changes between sessions.

Two real-data longitudinal studies support the premise that temporal history carries predictive signal. In a 12-week cohort of 229 patients with major depressive disorder, adding dynamic longitudinal features raised the best overall trajectory-classification accuracy from below 50% to 60.87% [46]. In a separate year-long study, initial and intermediate PHQ-9 states were the most important predictors of later deterioration [22]. Direct evidence for the memory module itself awaits a real multi-session corpus with session-level PHQ supervision (Appendix I).

G.4 Performance by Trajectory Type

To examine performance by trajectory type, we group the 738 test trajectories into the benchmark’s prespecified patterns of PHQ-8 change: improving (93), worsening (58), stable (358), and fluctuating (229). EMOTRACK’s advantage on LC8-QWEN35-35B is concentrated on stable trajectories, where it has the lowest MAE of any method, while both Lau et al. variants lead on the three non-stable subgroups (Table 17). The grouping uses the benchmark’s 5-point endpoint-change criterion; “stable” still admits session-to-session variation below 5 points. On the stable subgroup, the largest, EMOTRACK reaches 2.1152 against 2.2455 for the best baseline. It stays competitive on improving (2.9534 against 2.9167) and fluctuating (3.0760 against 2.8740) trajectories and is 0.3224 behind the best method on worsening cases (3.4494 against 3.1270). This matches the aggregate LC8-QWEN35-35B result in Table 8 and locates the residual gap in trajectories that change.

G.5 Runtime and Inference Cost

EMOTRACK’s inference cost is far below that of the neural baselines that re-encode transcripts: 0.029 single-GPU hours on LC8-QWEN35-35B against 7.23 for Lau et al. and 23.72 for its Qwen3-embedding variant (Table 18). We time inference on one GPU per run after feature caching, and the measurements exclude feature extraction and training; they therefore represent inference-stage computation rather than complete end-to-end runtime. On DAIC-WOZ the same pattern holds at a smaller scale (0.004 hours for EMOTRACK against 0.254 and 0.636 for the two Lau et al. variants), and the prompted methods and EnsemBERT, which reuse cached features, are similarly cheap. The EMOTRACK predictor therefore adds little inference cost on top of feature extraction.

Table 17: LC8-QWEN35-35B test MAE stratified by the trajectory types defined by the benchmark [19]. Bold marks the lowest mean in each column.

Method	Overall MAE ↓			
	Improving ($n=93$)	Worsening ($n=58$)	Stable ($n=358$)	Fluctuating ($n=229$)
AIDA (now)	3.3936 ± 0.0128	3.7111 ± 0.0086	3.0385 ± 0.0100	3.3121 ± 0.0059
AIDA (all)	3.7581 ± 0.0095	3.8885 ± 0.0147	2.9432 ± 0.0035	3.3672 ± 0.0170
LMIQ (now)	3.1859 ± 0.0291	3.5164 ± 0.0301	2.6911 ± 0.0170	3.2094 ± 0.0219
LMIQ (all)	3.5495 ± 0.0895	3.7239 ± 0.0404	2.4882 ± 0.0186	3.2599 ± 0.0242
Milintsevich et al.	3.0254 ± 0.0660	3.2949 ± 0.0941	2.3030 ± 0.0437	3.0227 ± 0.0781
Lau et al.	2.9893 ± 0.0647	3.1561 ± 0.1006	2.2555 ± 0.0633	2.9053 ± 0.0943
Lau et al. (Qwen3-Emb)	2.9167 ± 0.0720	3.1270 ± 0.0570	2.2455 ± 0.0586	2.8740 ± 0.0620
EMOTRACK	2.9534 ± 0.1090	3.4494 ± 0.0635	2.1152 ± 0.0680	3.0760 ± 0.0917

Table 18: Inference time in single-GPU hours. Values are mean with standard deviation.

Method	Inference	
	LC8-QWEN35-35B	DAIC-WOZ
AIDA (now)	0.000±0.000	0.000±0.000
LMIQ (now)	0.011±0.000	0.000±0.000
Milintsevich et al.	5.74±0.48	0.330±0.003
EnsemBERT	0.004±0.000	0.001±0.000
EnsemBERT (Qwen3-Emb)	0.017±0.001	0.002±0.000
Lau et al.	7.23±3.90	0.254±0.010
Lau et al. (Qwen3-Emb)	23.72±12.78	0.636±0.003
EMOTRACK (ours)	0.029±0.001	0.004±0.000

H Asset Licenses and Intended Use

Table 19 lists the assets used for evaluation or upstream representations; the benchmarks are used under their release terms and are not redistributed.

Table 19: Assets used in this paper.

Asset	Use	License/access terms
LongCounsel-8 (LC8-QWEN35-35B, LC8-GPT54-MINI, LC8-GPT56- LUNA) [19]	Synthetic longitudinal PHQ-8 benchmark.	Obtained from the benchmark release and used under its stated terms; not redistributed.
DAIC-WOZ and E-DAIC [9, 33]	Real single-session PHQ-8 benchmark and its frozen-transfer set.	USC ICT application-based access for academic/non-profit research; used under provider terms and not redistributed. https://dcapswoz.ict.usc.edu/
Qwen models [31]	Frozen turn embeddings and AIDA-style feature extraction.	Apache License 2.0 as listed on the model cards.
Baseline methods [16, 17, 25, 32, 34]	Reimplemented or adapted comparisons.	Original papers are cited; external implementations are used under their respective terms and are not redistributed.

Intended use and safeguards. EMOTRACK is a research model for transcript-based PHQ-8 estimation, not diagnosis, autonomous triage, or intervention. Releases should document limitations, avoid redistributing restricted transcripts, and require human review. Users must obtain DAIC-WOZ, E-DAIC, and LongCounsel-8 separately and follow the providers’ access and citation requirements.

I Limitations

Real longitudinal clinical validation. Although LongCounsel-8 enables controlled evaluation of longitudinal PHQ-8 tracking and DAIC-WOZ provides a real-data anchor, real multi-session counseling datasets with standardized session-level PHQ supervision remain limited. Therefore, our results should be interpreted as benchmark evidence for robust transcript-based depression tracking, rather than as clinical validation. We plan to extend the evaluation to real longitudinal patient data in collaboration with clinical partners.

Open-source upstream models. EMOTRACK uses open-source upstream models for utterance embedding and clinical-feature extraction. This design improves accessibility and reproducibility, but may not reflect the strongest performance attainable with frontier closed-source models. The extractor swap in Appendix F.7 shows that the advantage over the matched AIDA regressor holds under a proprietary extractor on DAIC-WOZ; a wider study of upstream models, including their effect on calibration and on the longitudinal sets, remains future work.

Total score versus individual PHQ-8 items. Our claims concern the PHQ-8 total score. The PHQ-8 item query readout provides an architectural prior for total-score prediction. Auxiliary item-level supervision improves total-score accuracy on LC8-QWEN35-35B but is disabled in the selected DAIC-WOZ configuration; per-item errors are reported as diagnostics for the readout. At the item-level error magnitudes observed here, individual item predictions are not accurate enough to serve as symptom-level assessments. Establishing item-level reliability would require larger real datasets with per-item supervision together with item-wise calibration and agreement analysis, which we leave to future work.

PHQ-8 prediction versus clinical decision support. Our evaluation focuses on PHQ-8 score prediction. PHQ-8 is a widely used and standardized depression-severity measure, making it an appropriate target for controlled comparison; however, PHQ-8 prediction covers only part of clinical decision support. Practical deployment would also require calibration, uncertainty estimation, threshold-based triage beyond the false-negative penalty studied here, privacy safeguards, and human-in-the-loop review. EMOTRACK should therefore be viewed as a depression-tracking component rather than an autonomous diagnostic or intervention system.