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 is widely adopted across online mental health platforms, where automatically estimating depressive symptom severity from transcripts is vital for client safety and clinical monitoring. We therefore study prediction of the standard eight-item Patient Health Questionnaire (PHQ-8) total score directly from counseling sessions. Existing clinical-based methods obtain structured assessments through large language model inference but summarize away individual client turns. Training-based methods preserve turn semantics but must learn clinical structure from limited labels. Integrating these representations requires preserving clinical-feature identity and selecting evidence relevant to each symptom. We propose **EmoTrack**, which jointly encodes clinical-feature tokens and frozen client-turn embeddings, then uses PHQ-8 item queries to estimate total severity. Optional gated memory incorporates evidence from the preceding session. To support controlled evaluation in the multi-session setting, we introduce COUNSELARC, a synthetic counseling dataset with session-level PHQ-8 supervision for predicting current-session severity with optional prior-session context. Across real DAIC-WOZ interviews, transfer to E-DAIC, and three longitudinal datasets, EmoTrack achieves the lowest normalized aggregate error. It reduces DAIC-WOZ mean absolute error by 12.5% and aggregate error by approximately 8.6%, each relative to the strongest baseline for that measure.

1 Introduction

Text-based counseling has become increasingly prevalent as the primary mode of care across online mental health platforms [18, 26, 39]. On these platforms, automatically estimating depressive symptom severity directly from session transcripts is critical for ensuring client safety, enabling timely risk detection and escalation. To ground this task in established clinical practice, we study the prediction of the widely adopted eight-item Patient Health Questionnaire (PHQ-8; scored 0–24) [12] from counseling transcripts. Furthermore, because online counseling often spans multiple encounters, earlier sessions can provide valuable longitudinal context for assessing a client's current state.

Existing methods obtain clinical information through two complementary routes. *Clinical-based methods* rely mainly on LLM inference to complete questionnaires or score clinician-style features, which are then mapped to severity estimates [15, 31]. These assessments supply explicit clinical structure but summarize the session without retaining individual client turns as separate predictor inputs. *Training-based methods* train predictors directly on transcripts or their semantic embeddings [14, 22, 29]. They preserve local conversational detail but must learn its relationship to symptom severity from scarce labeled transcripts. Combining prompted clinical features with turn-level semantics can bring clinical structure into transcript modeling under limited supervision.

Code: github.com/Xtra-Computing/EmoTrack Dataset: huggingface.co/datasets/hiddensev/LongCounsel-8

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This combination presents two challenges for single-session estimation and a further challenge for longitudinal use. **(i) Clinical scores and dialogue turns have mismatched structures.** Assessments form a small set of named scalars, whereas client turns form a variable-length sequence of text embeddings. Fusing them must retain what each score represents while connecting it to conversational evidence. **(ii) Different PHQ-8 items require different evidence.** A turn about insomnia may inform the sleep item, whereas feeling worthless may inform low self-worth. A pooled session vector forces all item predictors to rely on the same compressed evidence. **(iii) History may be missing or outdated.** Prior disclosures can clarify indirect statements but may mislead prediction after symptoms change. First or isolated sessions have no history.

To address these challenges, we build on the key idea of representing clinical assessments and client turns as distinct, retrievable evidence for each PHQ-8 item, and propose **EmoTrack** with three corresponding components. (i) To align these mismatched structures, we convert both inputs into a shared token space: clinical-feature tokens combine a score projection with feature- and group-identity embeddings, while client-turn tokens project frozen semantic embeddings for joint Transformer encoding. (ii) Instead of pooling a single session vector, we assign one learned query per PHQ-8 item to retrieve symptom-relevant evidence, whose resulting item predictions sum to the total score. (iii) To handle missing or shifting history, an optional memory module compresses the preceding session via symptom-specific retrieval and regulates its contribution with learned gates, leaving the current-session architecture intact when history is absent.

Beyond modeling, evaluating longitudinal symptom tracking faces a severe data bottleneck: it requires multi-session counseling dialogues paired with session-level symptom targets. However, real clinical interviews such as DAIC-WOZ only provide single-session PHQ-8 supervision [8], while existing multi-session counseling corpora focus on dialogue quality or continuity without aligned clinical labels [24, 26]. To bridge this evaluation gap, we introduce COUNSELARC, a synthetic longitudinal benchmark that grounds five-session counseling dialogues in client profiles, empirical severity trajectories, symptom compositions, and private behavioral cues, thoroughly audited for both label fidelity and dialogue realism.

Our contributions are threefold. (1) We introduce EmoTrack for joint clinical–semantic PHQ-8 estimation, with item-specific queries and optional previous-session memory. (2) We introduce COUNSELARC, a controlled synthetic multi-session counseling dataset with prescribed symptom trajectories and session-level PHQ-8 supervision, supporting current-session severity prediction under partial symptom disclosure and cross-session continuity. (3) We evaluate on real DAIC-WOZ interviews, frozen transfer to E-DAIC, and three longitudinal datasets, with controlled representation, readout, objective, and memory ablations. EmoTrack reduces DAIC-WOZ MAE by 12.5% relative to AIDA and achieves the lowest normalized aggregate error across the five datasets.

2 Related work

We organize transcript-based depression severity prediction by how clinical evidence enters the model. Clinical-based methods train downstream regressors on prompted assessments, while training-based methods learn from transcripts or turn embeddings.

Table 1: Methods for text-based depression severity prediction. Clinical knowledge denotes prompted clinical features, and training denotes learning from transcripts or turn embeddings.

Study	Clinical knowledge	Training	Longitudinal	Target scale
Lau et al. [14]	✗	✓	✗	PHQ-8
Milintsevich et al. [22]	✗	✓	✗	PHQ-8
PsychiatricPrism [2]	✗	✓	✗	PHQ-8
EnsemBERT [29]	✗	✓	✗	BDI-II
LMIQ [31]	✓	✗	✗	PHQ-8
AIDA [15]	✓	✗	✗	PHQ-8
EmoTrack (ours)	✓	✓	✓	PHQ-8

Clinical-based methods. Prompted LLMs convert transcripts into structured questionnaire responses or clinician-style assessments. LMIQ completes questionnaires and supplies the coded

responses to a random-forest regressor [31]. AIDA scores 23 clinical, linguistic, and cognitive features and maps them to PHQ-8 severity with regression [15]. These methods provide explicit clinical structure, but their session-level feature vectors do not retain individual client turns as separate predictor inputs. EmoTrack jointly represents prompted scores and client turns, enabling each PHQ-8 item query to access both sources of evidence.

Training-based methods. Supervised predictors learn from transcripts or turn embeddings under limited labeled data. Lau et al. [14] use parameter-efficient text-encoder tuning. Milintsevich et al. [22] predict PHQ-8 items hierarchically and sum their scores. PsychiatricPrism [2] uses a transformer-based symptom predictor and studies psychiatrist annotations. EnsemBERT [29] predicts BDI-II items from social-media text and is adapted here to PHQ-8. These methods retain local text evidence, while EmoTrack additionally supplies prompted clinical-feature tokens and optional previous-session memory (Table 1). Because interviewer prompts can provide shortcuts [4], we focus on client turns and test speaker-view alternatives.

Counseling data and longitudinal evaluation. DAIC-WOZ and E-DAIC provide real single-interview evaluation [8, 30]. Counseling resources such as PsyDial, CACTUS, and PsychEval broaden dialogue coverage [16, 24, 26], but do not provide the controlled five-session PHQ-8 trajectories needed here. We therefore introduce COUNSELARC, pairing synthetic counseling trajectories with per-session symptom states and simulated self-reports. Section 5 describes its construction, and Appendix A compares the wider method and dataset landscape. This resource complements real single-session evaluation with controlled longitudinal tests.

3 Problem formulation

We estimate depression severity from the client's language in a counseling transcript. A *turn* is one utterance, and a *session* is one counseling encounter. Single-session estimation uses the current transcript, while longitudinal estimation may also use earlier sessions from the same client. The target is the current session's PHQ-8 score.

Clinical target. The eight PHQ-8 symptoms [12] are anhedonia, depressed mood, sleep disturbance, fatigue, appetite disturbance, low self-worth or guilt, concentration difficulty, and psychomotor change. Questionnaire responses are ordinal, with item values in $\{0, 1, 2, 3\}$ and total $Y = \sum_{j=1}^{J} y_j \in [0, 24]$, where $J = 8$. The synthetic benchmark also uses fractional hidden states and averaged self-reports within these bounds (Appendix C).

Prediction task. Given the current transcript $\mathcal{T}^{(t)}$ and available history $\mathcal{H}^{(t-1)} = \{\mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(t-1)}\}$, we learn

$$f_\theta\big(\mathcal{T}^{(t)}, \mathcal{H}^{(t-1)}\big) \to \hat{\mathbf{y}}^{(t)} \in [0, 3]^J, \qquad \hat{Y}^{(t)} = \sum_{j=1}^{J} \hat{y}_j^{(t)}. \tag{1}$$

History is empty for the first or only session. EmoTrack uses one previous session by default, with two-session and full-history inputs evaluated in Appendix E.7. Past questionnaire labels are not model inputs.

4 Approach: EmoTrack

EmoTrack combines structured clinical features with client-turn semantics to estimate PHQ-8 severity. The two input branches form a shared token sequence, and learned PHQ-8 item queries retrieve symptom-specific evidence from an encoder–decoder backbone. An optional memory module incorporates previous-session context before the prediction head. Figure 1 summarizes the architecture, and Appendix B details its forward computation.

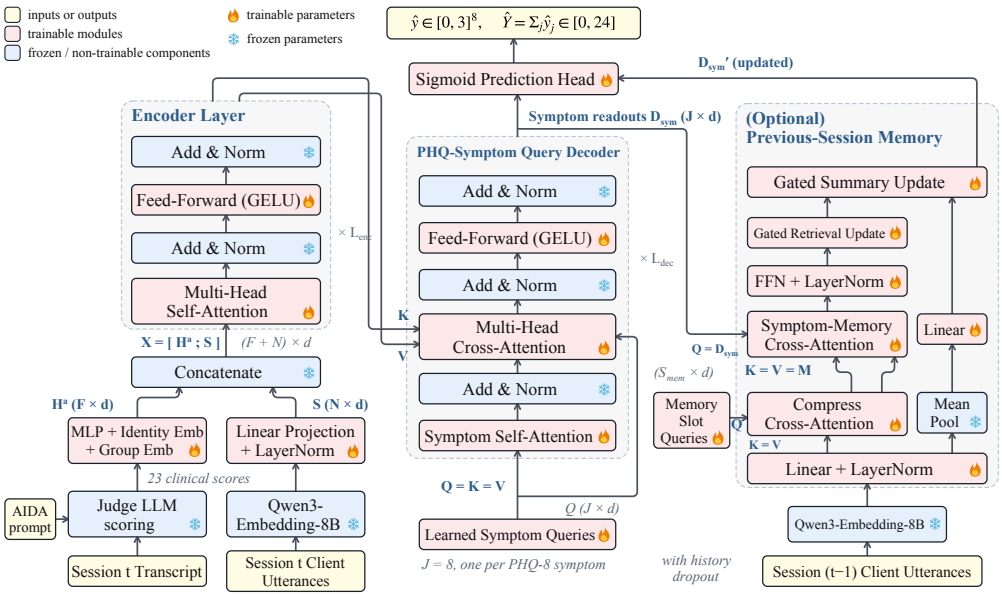


Figure 1: EmoTrack architecture for PHQ-8 severity estimation with optional previous-session memory.

4.1 Clinical and semantic inputs

The clinical branch provides explicit symptom-related evidence, while the dialogue branch retains semantic information that a fixed feature schema may omit. Both use the current client-side transcript $\mathcal{T}^{(t)}$. The input and speaker ablations in Section 6.4 test these choices.

Clinical-feature tokens. Following AIDA [15], we prompt an LLM to score $F = 23$ features comprising eight PHQ-aligned indicators, five linguistic-pattern features, and ten cognitive-distortion markers. Each score lies in $[0, 10]$ and is standardized using training-split statistics. For standardized score z_i, the token is

$$\mathbf{h}_i = \mathrm{MLP}_{\mathrm{score}}(z_i) + \mathbf{e}_i^{\mathrm{feat}} + \mathbf{e}_{g(i)}^{\mathrm{group}}. \tag{2}$$

The learned feature and group embeddings identify the score and its clinical group $g(i) \in \{1, 2, 3\}$. Stacking the tokens gives $\mathbf{H}^{\mathrm{c}} \in \mathbb{R}^{F \times d}$.

Client-turn tokens. A frozen sentence encoder maps each client turn to $\mathbf{u}_n \in \mathbb{R}^{d_e}$, and a trainable projection gives $\mathbf{s}_n = \mathrm{LayerNorm}(W_{\mathrm{proj}}\mathbf{u}_n + \mathbf{b}_{\mathrm{proj}})$. We remove counselor and empty turns, retain the first $N_{\max}$ client turns in chronological order, and encode each client turn as one fixed-size vector. The resulting sequence is $\mathbf{S} \in \mathbb{R}^{N \times d}$. Validation on DAIC-WOZ selects $N_{\max} = 80$, with cap sensitivity shown in Figure 5. The longitudinal sessions contain 20 client turns and fit within this cap.

4.2 Encoder–decoder and PHQ-8 query readout

Self-attention lets clinical and semantic tokens exchange information before prediction. The encoder processes $\mathbf{X} = [\mathbf{H}^{\mathrm{c}}; \mathbf{S}] \in \mathbb{R}^{(F+N) \times d}$ with a padding mask that excludes padded turns from attention. The selected backbone uses two encoder layers (three on DAIC-WOZ) and four decoder layers, with depth ablations reported in Appendix E.2.

Item-specific readout allows each symptom query to retrieve relevant evidence. Eight learned query vectors $\mathbf{Q} \in \mathbb{R}^{J \times d}$, one per PHQ-8 item, cross-attend to the encoded tokens. The decoder returns $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_J]^\top \in \mathbb{R}^{J \times d}$. Table 4 compares this readout with mean pooling, and Table 3 tests standard regressors on the same upstream inputs.

4.3 Prediction and training objective

After any memory refinement, the prediction head maps each item representation to $\hat{y}_j = 3\sigma(\mathbf{w}_j^\top \mathbf{d}_j + b_j) \in [0, 3]$, and sums these values into $\hat{Y} = \sum_j \hat{y}_j$. The training objective combines a total-score Huber loss with optional item-level supervision,

$$\mathcal{L} = \underbrace{\mathrm{Huber}(\hat{Y}, Y)}_{\mathcal{L}_{\mathrm{agg}}} + \lambda_{\mathrm{sym}} \underbrace{\frac{1}{J}\sum_{j=1}^{J} \mathrm{Huber}(\hat{y}_j, y_j)}_{\mathcal{L}_{\mathrm{sym}}}. \tag{3}$$

The auxiliary term follows symptom-resolved severity estimation [2, 22]. Validation selects $\lambda_{\mathrm{sym}} = 0.5$ on the main COUNSELARC set and 0 on DAIC-WOZ and the complementary dataset. Thus, the same architecture supports aggregate-only training. We use AdamW [19], learning-rate warmup and decay, gradient clipping, and early stopping on validation loss (Appendix D).

Optional false-negative penalty. When missing a screen-positive session is a separate evaluation concern, training can add

$$\mathcal{L}_{\mathrm{FN}} = \lambda_{\mathrm{FN}} \mathbf{1}[Y \geq 10] \max(0, 10 - \hat{Y}) \tag{4}$$

to Eq. (3). The term penalizes a prediction below the PHQ-8 screening threshold of 10 for a reference-positive session [12]. All default comparisons use $\lambda_{\mathrm{FN}} = 0$. A separate validation-FNR selection evaluates this option in Section 6.3.

4.4 Optional previous-session memory

For longitudinal inputs, memory refines the current-session item representations using earlier client turns. The default uses the immediately previous session and bypasses memory when history is absent. This keeps the single-session prediction pathway intact. Appendix E.7 evaluates longer histories with the same compressor.

Compressed retrieval. We project previous-session embeddings into $\mathbf{V}^{(t-1)} \in \mathbb{R}^{M \times d}$. Learned slot queries $\mathbf{P} \in \mathbb{R}^{S \times d}$ compress these tokens into fixed-size memory,

$$\mathbf{M} = \mathrm{LN}\big(\mathrm{MHA}(\mathbf{P}, \mathbf{V}^{(t-1)}, \mathbf{V}^{(t-1)})\big) \in \mathbb{R}^{S \times d}. \tag{5}$$

Here, MHA denotes multi-head attention and LN denotes layer normalization. Current item representations attend to $\mathbf{M}$, producing retrieved evidence $\mathbf{C} \in \mathbb{R}^{J \times d}$. A gated residual update controls its contribution,

$$\mathbf{D} \leftarrow \mathbf{D} + \sigma\big(W_g[\mathbf{D}; \mathbf{C}]\big) \odot W_o \mathbf{C}. \tag{6}$$

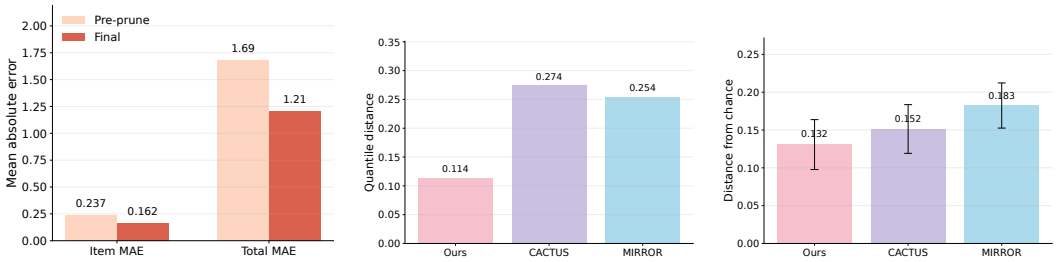
Summary and regularization. A second path mean-pools the projected previous-session tokens into $\bar{\mathbf{v}} \in \mathbb{R}^d$ and broadcasts this vector to all items,

$$\mathbf{D} \leftarrow \mathbf{D} + \sigma\big(W_{g'}[\mathbf{D}; \bar{\mathbf{v}}]\big) \odot \bar{\mathbf{v}}. \tag{7}$$

Validation selects the combined retrieval and summary paths, with four slots on COUNSELARC and 16 on the complementary dataset. During training, history is dropped with probability p_{hist} to limit dependence on prior context. The selected rates are 0.1 and 0.2, respectively. Figure 5 shows the dropout sweep. Appendix E.5 and Appendix E.6 separate path and capacity effects.

5 COUNSELARC: dataset construction

We introduce COUNSELARC and its construction protocol for controlled current-session PHQ-8 prediction with optional history. Empirical severity trajectories define private symptom states, which guide counseling dialogue and provide a reference for imperfect self-report labels.



(a) Self-report error before and after cue pruning.

(b) Psycholinguistic distance.

(c) Judge distance from chance.

Figure 2: Validation of COUNSELARC ("Ours") through self-report fidelity and dialogue similarity to RealCBT. Lower values indicate smaller errors or closer alignment with RealCBT.

5.1 Construction

Profiles and symptom trajectories. We first convert PsychEval-derived cases [24] into stable client profiles and session plans, preserving backgrounds and presenting concerns while rewriting session themes as client-side disclosure beats. We then reconstruct five-visit severity trajectories from PSYCHE-D quarterly observations [20]. For each PHQ-9 total-score anchor, we select a complete NHANES DPQ response vector with the matching total [5]. Retaining its first eight items produces the hidden PHQ-8 symptom vector, whose sum defines the session's hidden PHQ-8 total. Selecting complete response vectors preserves observed symptom compositions.

Private cues and counseling dialogue. Private cues determine symptom content, while RealCBT-derived surface constraints shape how it is expressed. Each symptom score is translated into private behavioral cues describing lived experience without questionnaire wording or numeric labels. We prune candidate cues through joint composition checks and sample five cues per symptom during generation. The client simulator receives these cues, its profile, and cross-session memory. The counselor receives public dialogue context and session guidance. RealCBT [35] calibrates speaker- and stage-specific turn length, question form, discourse openings, and lexical overlap.

Self-report labels and partitions. After each session, the simulated client completes five PHQ-8 passes with shuffled symptom and option orders, and we average the responses itemwise. Models train on these self-reports and are evaluated against the hidden generating state, whose numeric scores are excluded from dialogue. All five visits of a trajectory remain in one data split. Appendix C.1 details construction and filtering, and Appendices C.2–C.3 report the audits.

5.2 Dataset validation

The longitudinal benchmark requires reliable simulated labels and dialogue that resembles real counseling. Figure 2 evaluates these properties through three complementary audits, with the full analysis in Appendices C.2–C.3.

Self-report fidelity. Cue pruning substantially improves agreement between simulated self-reports and prescribed symptom states. Figure 2a reports item-level and total-score MAE between simulated self-reports and the hidden PHQ-8 vector when sampling five behavioral cues per symptom. Pruning reduces item MAE by 31.6% and total-score MAE by 28.4% while consistently increasing exact-match rates across item, profile, and total-score levels. Appendix Figure 6 provides the full exact-match and per-symptom breakdown.

Dialogue realism. COUNSELARC aligns more closely with real counseling transcripts in RealCBT than CACTUS and MIRROR across both linguistic and judge-based realism measures [11, 16, 35]. Figure 2b measures aggregate psycholinguistic quantile distance over speaker-specific turn length, readability, and affective statistics, while Figure 2c reports pointwise judge distance from chance to account for systematic judgment inversion. COUNSELARC more than halves the psycholinguistic distance relative to either baseline and reduces judge distance from chance by 13.2% relative to

CACTUS and 27.9% relative to MIRROR. Appendix Figure 7c details the feature-level audit across all seven psycholinguistic statistics.

6 Experiments

We evaluate current-session PHQ-8 prediction across real single-session interviews, zero-shot protocol transfer, and multi-session longitudinal counseling, and further analyze per-symptom error and directional bias (Appendix E.11), session-to-session transitions (Appendix E.12), trajectory-type stratification (Appendix E.13), optional false-negative screening, and component ablations.

6.1 Experimental settings

Datasets and protocols. We evaluate on five datasets. DAIC-WOZ [8] provides 189 real PHQ-8-labeled interviews, with 107/35/47 training/validation/test participants. For E-DAIC [30], models trained and selected on DAIC-WOZ are evaluated without adaptation. For multi-session evaluation, we use COUNSELARC, generated with Qwen3.5-35B-A3B. LongCounsel-8 [17] extends COUNSELARC with additional dialogue generators. We use two datasets from this work, LongCounsel-8-Luna and LongCounsel-8-GPT54-Mini (the complementary set), for cross-generator evaluation. COUNSELARC and LongCounsel-8-Luna each contain 3,690 trajectories (18,450 sessions), with the latter generated using GPT-5.6 Luna. LongCounsel-8-GPT54-Mini uses GPT-5.4-mini and contains 369 trajectories (1,845 sessions). All use 70/10/20% trajectory-level splits, with five sessions per trajectory and 20 client plus 20 counselor turns per session. Models train on simulated self-reports and are scored against hidden generating states. History contains earlier transcripts without past labels. The GPT-5.4-mini set differs in generator and scale.

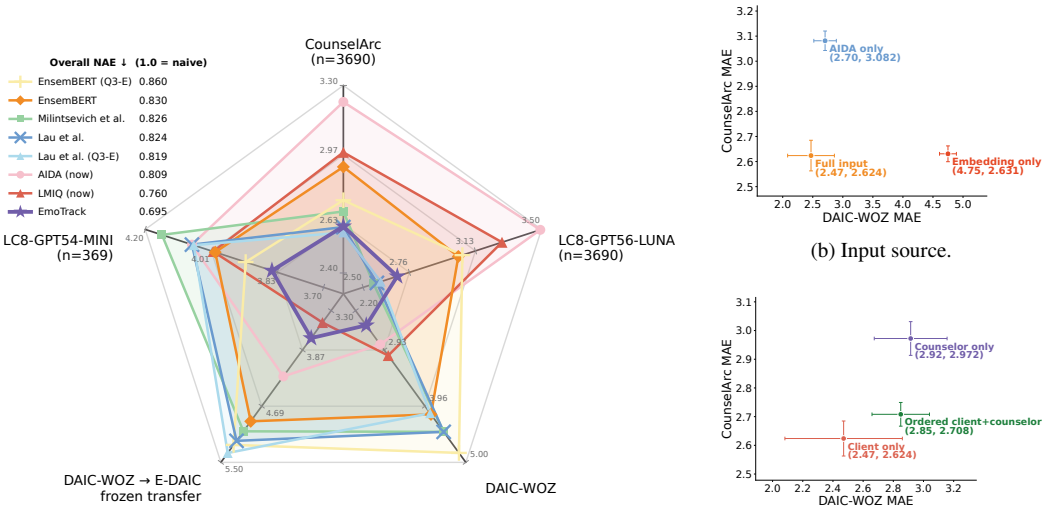
Model inputs and training. The frozen Qwen3-Embedding-8B encoder [43] produces a 4,096-dimensional vector per client turn. Qwen3.5-35B-A3B [28] extracts the 23 AIDA clinical scores. Each extraction seed is paired across downstream methods. We train with AdamW, batch size 32, learning rate 10^{-3}, weight decay 10^{-2}, 5% warmup followed by linear decay, gradient clipping at norm 1.0, and early stopping with patience 8 on validation loss. Appendix D gives dataset-specific configurations and hardware.

Metrics and selection. We report current-session PHQ-8 test MAE and screening rates as five-seed means, with error bars and $\pm$ values denoting one sample standard deviation across seeds. Transition and trajectory-type analyses stratify current-session score error. We select hyperparameters by validation MAE and apply early stopping on validation loss. Normalized aggregate error summarizes total-score accuracy across the five sets,

$$\mathrm{NAE} = \sqrt{\frac{1}{5}\sum_{k=1}^{5}\left(\frac{\mathrm{MAE}_k}{U_k}\right)^2}, \tag{8}$$

where U_k is the test MAE of a constant predictor that outputs the training-label median for set k. E-DAIC transfer retains the DAIC-WOZ training median. Each set has equal weight, lower is better, and one denotes the constant predictor. The optional screening analysis also reports false-negative rate (FNR) among reference-positive sessions and false-positive rate (FPR) among reference-negative sessions at a PHQ-8 threshold of 10. Its penalty weight is selected separately by validation FNR.

Baselines. For the clinical-based baselines, **AIDA** [15] fits ridge regression to 23 extracted features, and **LMIQ** [31] fits a random forest to LLM-completed questionnaire features. For the training-based baselines, **Lau et al.** [14] use a tuned RoBERTa question–response encoder and sentence-transformer branch. **Milintsevich et al.** [22] use hierarchical symptom prediction. **EnsemBERT** [29] combines symptom-relevant retrieval with attention, adapted here from BDI-II to PHQ-8. For AIDA and LMIQ, "now" uses the current session and "all" concatenates available history. Where applicable, trained baselines use both original encoders and Qwen3 embeddings. Comparisons share splits and extraction seeds.



(a) Cross-setting MAE.

(b) Input source.

(c) Speaker selection.

Figure 3: Total-score MAE across evaluation sets (a), input sources (b), and speaker selections (c). Point labels in (b,c) give (DAIC-WOZ MAE, COUNSELARC MAE).

6.2 Main results

Accuracy across settings. EmoTrack generalizes more effectively across single-session, transfer, and longitudinal settings than either clinical-based or training-based baselines alone. Figure 3a presents the cross-setting MAE alongside NAE, which measures overall cross-dataset accuracy. Whereas baselines struggle outside their primary regime, EmoTrack encloses the smallest radar footprint and achieves the lowest NAE, outperforming the strongest overall competitor, LMIQ (now), by approximately 8.6%. Appendix E.1 details the full per-dataset results.

Real single-session evaluation. Jointly modeling clinical features and client-turn semantics yields the most accurate severity predictions on real clinical interviews. Table 2 reports the five-seed test MAE on DAIC-WOZ, evaluating single-session estimation under scarce supervision. Whereas purely training-based predictors lag behind prompted clinical methods, EmoTrack achieves the lowest mean error and reduces MAE by 12.5% relative to the strongest baseline, AIDA (now).

Table 2: Total-score MAE on DAIC-WOZ, with the lowest mean in bold. Values are mean ± sample standard deviation over five seeds.

Metric	AIDA (now)	LMIQ (now)	Milintsevich et al.	EnsemBERT	EnsemBERT (Qwen3-Emb)	Lau et al.	Lau et al. (Qwen3-Emb)	**EmoTrack (Ours)**
MAE ↓	2.8234 ± 0.2099	3.0285 ± 0.0460	4.4370 ± 0.1197	4.1149 ± 0.3339	4.8255 ± 0.7277	4.4367 ± 0.5744	4.0927 ± 0.2795	**2.4708 ± 0.3892**

Transfer without adaptation. EmoTrack generalizes competitively to unseen interview protocols without fine-tuning, ranking second overall and substantially outperforming all training-based baselines. Figure 3a shows the frozen DAIC-WOZ-to-E-DAIC transfer MAE, measuring out-of-distribution accuracy when models trained on DAIC-WOZ are evaluated directly on E-DAIC. EmoTrack ranks second overall, trailing only LMIQ (now) by 6.5% while reducing MAE by 13.1% relative to AIDA (now) and by at least 24.6% relative to every training-based baseline. Appendix E.1 details the full results.

Longitudinal evaluation. EmoTrack maintains strong performance across diverse synthetic dialogue generators and achieves the lowest error under reduced supervision. Figure 3a displays the multi-session test MAE across the three longitudinal benchmarks, evaluating current-session estimation with prior-session context. On the smaller complementary set (LongCounsel-8-GPT54-Mini),

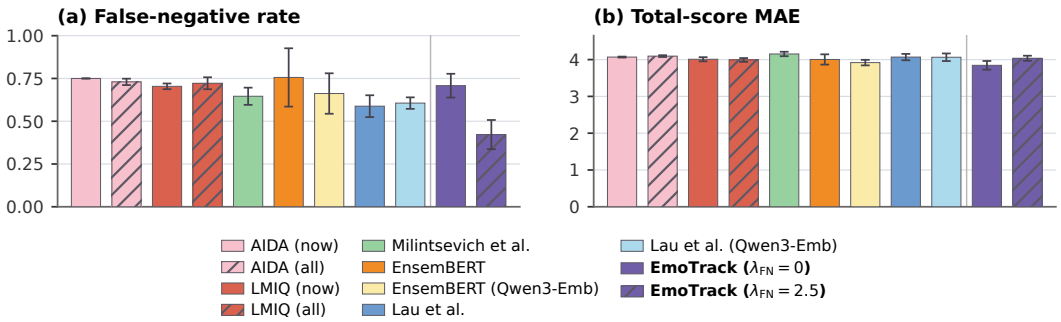


Figure 4: False-negative rate (a) and total-score MAE (b) on the complementary set for the baselines and EmoTrack with and without the optional penalty.

EmoTrack ranks first and reduces MAE by 1.9% relative to the strongest baseline, EnsemBERT (Qwen3-Emb), while staying competitive to the leading trained predictors on the two full-scale corpora. Appendix E.1 reports the complete longitudinal comparison.

6.3 Optional false-negative penalty

Applying the optional false-negative penalty markedly improves screening sensitivity for reference-positive sessions while introducing a controlled trade-off in score error. Figure 4 presents the false-negative rate (FNR) and total-score MAE across all methods on the complementary set. Relative to the default configuration ($\lambda_{\mathrm{FN}} = 0$), validation-selected $\lambda_{\mathrm{FN}} = 2.5$ lowers FNR by 40.4%—surpassing all baselines including the strongest competitor, Lau et al.—while increasing MAE by 5.0%. Appendix Table 16 reports the accompanying FPR trade-off, and Appendix E.10 provides the weight-sensitivity analysis.

6.4 Ablation study

We ablate input sources, speaker selection, predictor architecture, PHQ-symptom query decoding, history dropout, retained client turns, encoder–decoder depth (Appendix E.2), previous-session memory pathways (Appendix E.5), memory slot capacity (Appendix E.6), historical session scope (Appendix E.7), symptom-wise loss weighting (Appendix E.8), clinical-feature extractors (Appendix E.9), and inference runtime (Appendix E.14).

Input sources and speaker selection. The relative contribution of clinical scores and client-turn embeddings shifts with supervision scale, whereas client-only speech is consistently optimal across regimes. Figure 3b compares test MAE on DAIC-WOZ and COUNSELARC when ablating each input branch, and Figure 3c evaluates alternative speaker selections. On DAIC-WOZ, removing clinical features nearly doubles MAE while removing turn embeddings has a modest effect; conversely, on COUNSELARC, turn embeddings alone nearly match the full model whereas clinical features alone perform worse. For speaker selection, using only client turns achieves the lowest mean error on both datasets.

Predictor architecture. EmoTrack's backbone extracts more predictive signal from the combined inputs than standard tabular and neural regressors. Table 3 compares overall test MAE on COUNSELARC and DAIC-WOZ when EmoTrack, ridge regression, an MLP, and gradient-boosted trees are trained on identical clinical scores and client-turn embeddings. EmoTrack outperforms all three baselines, reducing MAE by 5.9% on COUNSELARC and 12.0% on DAIC-WOZ relative to gradient-boosted trees, the strongest standard regressor.

History dropout and retained client turns. Moderate history dropout and bounded client-turn context improve generalization by limiting reliance on noisy or excessive dialogue history. Figure 5 plots total-score test MAE across history-dropout rates on the longitudinal sets and retained-client-turn caps on DAIC-WOZ. A dropout rate of 0.1 achieves the lowest error on COUNSELARC, reducing

Table 3: Predictor comparison using the same AIDA features and Qwen3 embeddings on COUNSELARC and DAIC-WOZ.

Method	**Overall MAE ↓**	
	COUNSELARC	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**

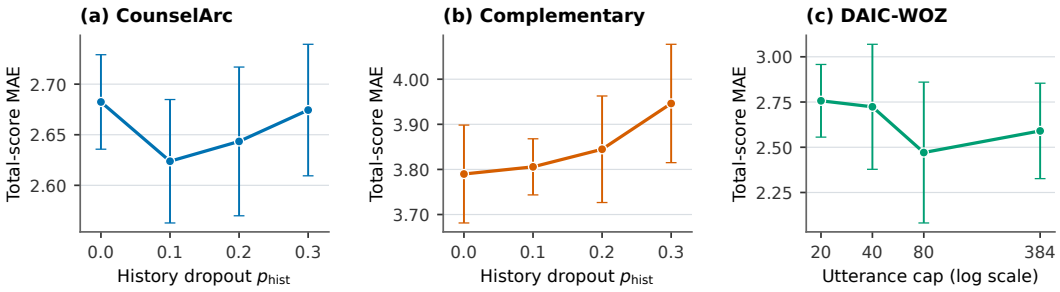


Figure 5: Total-score MAE by history-dropout rate on COUNSELARC (a) and the complementary set (b), and by retained-client-turn cap on DAIC-WOZ (c). Panel (c) uses a logarithmic horizontal axis.

MAE by 2.2% relative to no dropout, whereas the complementary set reaches its test minimum without dropout and selects 0.2 by validation (1.4% higher). On DAIC-WOZ, retaining 80 client turns outperforms both shorter and longer contexts, reducing MAE by 4.6% relative to 384 client turns and by at least 9.3% relative to 20 or 40 client turns while accommodating all 20-turn longitudinal sessions. Appendices E.3 and E.4 detail these sweeps.

PHQ-symptom query decoder. Decoding symptom-specific representations through learned PHQ-8 queries outperforms uniform mean pooling on the primary single-session and longitudinal benchmarks. Table 4 compares total-score test MAE between the PHQ-symptom query decoder and a mean-pooled encoder head across three datasets. Replacing the query decoder with mean pooling increases MAE by 4.3% on COUNSELARC and 8.8% on DAIC-WOZ, while the two readouts differ by only 0.5% on the complementary set.

Table 4: Effect of replacing the PHQ-symptom query decoder with a mean-pooled encoder head on COUNSELARC, DAIC-WOZ, and the complementary GPT-5.4-mini-generated dataset.

Prediction head	**Overall MAE ↓**		
	COUNSELARC	DAIC-WOZ	GPT-5.4-MINI COMP.
PHQ-symptom query decoder	**2.6238 ± 0.0610** †	**2.4708 ± 0.3892** †	3.8447 ± 0.1181 †
Mean-pooled encoder head	2.7364 ± 0.0411	2.6889 ± 0.1328	**3.8262 ± 0.1046**

7 Conclusion

EmoTrack bridges prompted clinical structure and client-turn semantics through identity-preserving joint encoding and PHQ-8 item queries, with an optional gated memory module for longitudinal use. We also introduce the COUNSELARC benchmark and its construction protocol for controlled session-level severity prediction, supported by comprehensive cue-fidelity and dialogue-similarity audits. Evaluation across five diverse datasets consistently yields the lowest normalized aggregate error, while ablations highlight dataset-dependent gains from historical memory and supervision.

Acknowledgments and Disclosure of Funding

We thank Adrian Toh, Clinical Psychologist, for his valuable clinical insights and expert guidance in clinical psychology, particularly in informing the design and development of the counselling dataset and depression severity estimation pipeline.

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority. This research is also in part supported by Zhaomin Wu's National Research Foundation (NRF) Postdoctoral Award.

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 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. doi: 10.18653/v1/2024.lrec-main.87. URL `https://aclanthology.org/2024.lrec-main.87/`.

[3] 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. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:20986–20995, 2025.

[4] Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, Adrian Lopez Monroy, and Petr Motlicek. Daic-woz: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 82–90, 2024. doi: 10.18653/v1/2024.clinicalnlp-1.8. URL `https://aclanthology.org/2024.clinicalnlp-1.8/`.

[5] 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`.

[6] 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 *North American Chapter of the Association for Computational Linguistics*, 2024.

[7] Cheryl R. Every-Wurtz. Counseling and psychotherapy transcripts, client narratives, and reference works. `https://search.alexanderstreet.com/psyc`, 2009. Commercial database, Alexander Street, part of Clarivate. Accessed 2026-04-02.

[8] 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 *International Conference on Language Resources and Evaluation*, 2014.

[9] Aylin Gunal, Bowen Yi, John Piette, Rada Mihalcea, and Ver'onica P'erez-Rosas. Examining spanish counseling with midas: a motivational interviewing dataset in spanish. *ArXiv*, abs/2502.08458, 2025.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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. doi: 10.3389/fpsyt.2023.1160291.

[15] Jae-Joong Lee, Jihoon Han, and Choong-Wan Woo. Interpretable depression assessment using a large language model. *PLOS Digital Health*, 5(2):e0001205, 2026. doi: 10.1371/journal.pdig.0001205. URL `https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0001205`.

[16] Suyeon Lee, Sunghwan Mac 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.

[17] 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.

[18] 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.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[20] 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.

[21] 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. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2021.

[22] Kirill Milintsevich, Kairit Sirts, and Gaël Dias. Towards automatic text-based estimation of depression through symptom prediction. *Brain informatics*, 10(1):4, 2023. doi: 10.1186/s40708-023-00185-9.

[23] 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.

[24] 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 and comprehensive ai psychological counselor. *arXiv preprint arXiv:2601.01802*, 2026.

[25] 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. *ArXiv*, abs/2506.01357, 2025.

[26] 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. doi: 10.18653/v1/2025.acl-long.1049. URL `https://aclanthology.org/2025.acl-long.1049/`.

[27] 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.

[28] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

[29] 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. doi: 10.1371/journal.pdig.0000848.

[30] 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.

[31] 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. doi: 10.18653/v1/2024.findings-emnlp.23. URL `https://aclanthology.org/2024.findings-emnlp.23/`.

[32] 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.

[33] Aseem Srivastava, Tharun Suresh, S. P. Lord, Md. Shad Akhtar, and Tanmoy Chakraborty. Counseling summarization using mental health knowledge guided utterance filtering. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[34] BN Suhas, Andrew M 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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[35] 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 *Conference on Empirical Methods in Natural Language Processing*, 2025.

[36] Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. Explainable depression detection using large language models on social media data. *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, 2024.

[37] Zixiu "Alex" Wu, Simone Balloccu, Vivek Kumar (Ph.D), Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-mi: A dataset of expert-annotated counselling dialogues. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181, 2022.

[38] 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.

[39] 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. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2025.

[40] Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. Cpsycoun: 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.

[41] 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.

[42] 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.

[43] 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`.

[44] 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 **15**
- A.1 Transcript-based depression severity prediction . . . 15
- A.2 Counseling and motivational interviewing datasets . . . 16

B Algorithmic details of EmoTrack **16**

C Construction and evaluation of COUNSELARC **16**
- C.1 Construction . . . 17
- C.2 Self-report fidelity audit . . . 20
- C.3 RealCBT similarity audit . . . 20

D Additional experiment settings **20**

E Additional experiments **22**
- E.1 Complete method comparison . . . 22
- E.2 Ablation study: encoder–decoder layers . . . 23
- E.3 Ablation study: history dropout . . . 23
- E.4 Ablation study: retained client turns . . . 23
- E.5 Ablation study: memory mechanism . . . 23
- E.6 Ablation study: memory slots . . . 24
- E.7 Ablation study: history scope . . . 24
- E.8 Ablation study: symptom-wise loss . . . 24
- E.9 Clinical-feature extractor . . . 25
- E.10 False-negative penalty . . . 25
- E.11 Per-item error and directional bias . . . 26
- E.12 Current-score error after session-to-session transitions . . . 26
- E.13 Current-score error by trajectory type . . . 27
- E.14 Runtime and inference cost . . . 27

F Limitations **27**

G Asset licenses and data statement **28**

A Additional related work

A.1 Transcript-based depression severity prediction

Table 5 compares transcript-based score prediction with multimodal and classification approaches, identifying the input evidence and prediction target for each method.

LMIQ fits a random forest to prompted questionnaire responses [31], and AIDA fits a regression model to prompted clinical-feature scores [15]. PsychiatricPrism trains a transformer and evaluates symptom predictions against expert annotations [2]. EmoTrack combines clinically structured features with supervised symptom prediction and retains client turns as separate evidence tokens.

Interpretation of transcript results also depends on how the interview was conducted. Burdisso et al. [4] identify interviewer-prompt shortcuts in DAIC-WOZ depression classification, motivating explicit checks of speaker selection. Milintsevich et al. [22] further observe that some symptoms are sparsely expressed in the transcript and that prediction errors vary by symptom. These findings motivate our speaker-view and per-symptom analyses.

Table 5: Approaches to depression evaluation.

Study	Modality	Data	Clinical Criteria	Output
Multimodal methods				
MoodCapture [23]	Image	Smartphone images	PHQ-8, DSM-5	Score
LLM+Landmarks [42]	Speech + Text	DAIC-WOZ	PHQ-8	Binary
Sadeghi et al. [32]	Video + Speech + Text	E-DAIC	PHQ-8	Score
SEGA [6]	Video + Speech + Text	DAIC-WOZ	PHQ-8	Score
Text-based methods with classification output				
DORIS [13]	Text	Weibo & Twitter posts	DSM-5	Binary
Wang et al. [36]	Text	Social-media posts	BDI	Binary
RED [41]	Text	DAIC-WOZ	PHQ-8	Binary
PsyGUARD [27]	Text	Counseling utterances	C-SSRS	Level
Text-based methods with score output				
Lau et al. [14]	Text	DAIC-WOZ	PHQ-8	Score
PsychiatricPrism [2]	Text	DAIC-WOZ	PHQ-8	Score
LMIQ [31]	Text	DAIC-WOZ	PHQ-8	Score
AIDA [15]	Text	DAIC-WOZ	PHQ-8	Score
Milintsevich et al. [22]	Text	DAIC-WOZ	PHQ-8	Score
EnsemBERT [29]	Text	Reddit posts	BDI-II	Score
EmoTrack (ours)	Text	DAIC-WOZ & COUNSELARC	PHQ-8	Score

A.2 Counseling and motivational interviewing datasets

Table 6 surveys counseling and motivational-interviewing resources by provenance, language, multi-session structure, and PHQ supervision. Resources based on human material include public demonstrations [21, 33, 35], trained-counselor role-play [9, 25, 37], reconstructed or rewritten conversations [18, 26], and licensed therapy transcripts [7]. The *real* provenance category includes role-play and reconstructed source material. Synthetic resources use different therapeutic frameworks and generation procedures [1, 3, 10, 11, 16, 24, 34, 38, 40].

These corpora support research on dialogue quality, therapeutic style, and conversational continuity. Within the resources summarized here, DAIC-WOZ [8] provides real single-session PHQ supervision, while COUNSELARC supplies synthetic multi-session targets prescribed during construction (Section 5). Section 6.1 specifies the resource versions used in our experiments.

B Algorithmic details of EmoTrack

Algorithm 1 summarizes the forward computation using the tokenizers, backbone, and memory module defined in Section 4.

C Construction and evaluation of COUNSELARC

COUNSELARC is constructed as a controlled multi-session counseling corpus for current-session PHQ-8 prediction with optional prior-session context. The construction follows a four-stage process that (1) defines a stable client profile and session plan, (2) assigns a five-visit depression trajectory, (3) decomposes each visit-level severity anchor into eight PHQ-8 symptom scores, and (4) renders those private symptom states as natural counseling dialogue. Empirical sources and deterministic validation rules guide each step.

Table 6: Counseling and motivational interviewing dialogue resources.

Work	Type	Lang.	Multi-session	PHQ-supervised
Real counseling / motivational interviewing resources				
KokoroChat [25]	Real	JA	✓	✗
HOPE [21]	Real	EN	✗	✗
MEMO [33]	Real	EN	✗	✗
Psych8K [18]	Real	EN	✗	✗
PsyDial-D0_m [26]	Real	ZH	✓	✗
Counseling and Psychotherapy Transcripts [7]	Real	EN	✓	✗
RealCBT [35]	Real	EN	✗	✗
AnnoMI [37]	Real	EN	✗	✗
MIDAS [9]	Real	ES	✗	✗
MentalChat16K [39]	Mixed	EN	✗	✗
DAIC-WOZ [8]	Real	EN	✗	✓
Synthetic counseling / motivational interviewing resources				
PsyDT / PsyDTCorpus [38]	Synthetic	ZH	✗	✗
PsyDial [26]	Synthetic	ZH	✓	✗
CACTUS [16]	Synthetic	EN	✗	✗
CPsyCoun / Memo2Demo [40]	Synthetic	ZH	✗	✗
HamRaz [1]	Synthetic	FA	✗	✗
MIRROR [11]	Synthetic	EN	✗	✗
KMI [10]	Synthetic	KO	✗	✗
Thousand Voices of Trauma [34]	Synthetic	EN	✗	✗
PsychEval [24]	Synthetic	ZH	✓	✗
COUNSELARC (ours)	Synthetic	EN	✓	✓

C.1 Construction

Generation chain. Each trajectory contains five counseling sessions for one client. For session t, the latent symptom state is a symptom-level PHQ-8 vector

$$\mathbf{y}^{(t)} = [y_1^{(t)}, \ldots, y_8^{(t)}], \qquad y_j^{(t)} \in [0, 3], \tag{9}$$

whose sum gives the visit-level PHQ-8 total score. PsychEval-derived cases [24] provide stable client backgrounds, presenting concerns, and session-stage structures. PSYCHE-D [20] provides the longitudinal severity anchors, aligning the five-visit changes with empirical depression trajectories. NHANES DPQ responses [5] then provide observed symptom co-occurrence patterns. For each anchor, we sample a complete PHQ-style response vector and retain the first eight symptoms as the latent PHQ-8 state.

The client-profile step deterministically converts PsychEval-derived cases. It preserves stable identity, biography, presenting concern, topic, growth history, beliefs, and target behaviors, while removing counselor-only administrative material and process instructions. Session themes are rewritten as client-side disclosure beats that specify a concrete event, bodily or contextual triggers, feared meaning, avoidance cost, and desired help. Profiles without valid client information or session plans are removed before trajectory assignment.

For trajectory reconstruction, PSYCHE-D quarterly observations are chained into five anchors. The first visit uses the first available start score, and subsequent visits use the corresponding quarter-end scores. When adjacent quarterly windows disagree on a shared boundary, the two boundary values are averaged and rounded. For each reconstructed PHQ-9 anchor, NHANES DPQ rows with complete symptom responses are grouped by total score. A deterministic hash of the global seed, participant id, visit index, and anchor score selects one matching response vector. The first eight PHQ symptoms form the PHQ-8 vector, and their sum becomes the latent PHQ-8 total. The decomposition uses

Algorithm 1 EmoTrack forward computation

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

Ensure: PHQ-8 symptom predictions $\hat{\mathbf{y}} \in [0,3]^J$, where $J = 8$

// Current-session representation

$\mathbf{a}^{(t)} \leftarrow \phi_{\mathrm{LLM}}(\mathcal{T}^{(t)})$

$\mathbf{H}^{\mathrm{c}} \leftarrow \Phi_{\mathrm{clin}}(\mathbf{a}^{(t)})$

$\mathbf{S}^{(t)} \leftarrow \Phi_{\mathrm{dlg}}\big(E(\mathcal{T}^{(t)})\big)$

$\mathbf{X}^{(t)} \leftarrow [\mathbf{H}^{\mathrm{c}}; \mathbf{S}^{(t)}]$

// Current-session PHQ-8 symptom encoding

$\mathbf{D}^{(t)} \leftarrow \mathcal{B}_\theta(\mathbf{X}^{(t)}, \mathbf{Q})$

// Optional previous-session memory refinement

if $\mathcal{T}^{(t-1)}$ is available and not dropped by history dropout **then**

  $\mathbf{V}^{(t-1)} \leftarrow \Phi_{\mathrm{hist}}\big(E(\mathcal{T}^{(t-1)})\big)$

  $\mathbf{D}^{(t)} \leftarrow \mathcal{R}_\theta\big(\mathbf{D}^{(t)}, \mathbf{V}^{(t-1)}\big)$

end if

// Symptom-level PHQ-8 prediction

$\hat{y}_j \leftarrow 3 \cdot \sigma(\mathbf{w}_j^\top \mathbf{d}_j^{(t)} + b_j), \quad j = 1, \ldots, J$

return $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_J]$

exact observed NHANES total-score buckets, with all-zero and all-maximum anchors handled as deterministic boundary cases.

Condensed prompt templates. The following templates are used only after deterministic profile and trajectory construction. The labels identify the generation stages and appear outside the prompt text. Bracketed spans denote instance-specific content instantiated from one client's profile, one session's private symptom cues, the public dialogue history, retrieved memories, and the current turn plan.

Behavior-cue generation.

> Given a PHQ-8 symptom, an ordinal severity level, and the intended direction when applicable, write short candidate behavioral cues that could naturally appear in a counseling client. Each cue must describe one lived-experience signal in 12–30 words. Do not mention PHQ, questionnaires, scores, symptom names, response options, or measurement language. Do not write first-person questionnaire answers or direct paraphrases of the PHQ symptom. Avoid cues that primarily express another PHQ-8 symptom.

Counselor-turn generation.

> You are the counselor. Given the public conversation so far, the session focus, a brief carryover note, relevant past-session memories, the target counseling move for this turn, and the RealCBT-derived turn form, write only the next counselor utterance in English. Follow the specified surface form, continue the session naturally, avoid repetition and premature closure, and do not mention latent states, PHQ, scores, or hidden generation instructions. Return only the visible counselor message in the required structured form.

Client-turn generation.

> You are the client described by the private profile. Given the public conversation so far, the current session disclosure beat, private symptom cues, stable background, carryover note, relevant memories, and the RealCBT-derived turn form, write only the next client utterance in English. Express the private cues indirectly through ordinary speech, partial disclosure, hesitation, affect, or concrete examples. Do not name PHQ, scores, questionnaire symptoms, cue lists, or hidden instructions; do not speak like a counselor. Return the visible client message and a short emotion phrase in the required structured form.

Self-report generation.

> After the session, answer PHQ-8 as the same client using the private symptom cues, stable profile, carryover note, and completed transcript. For each shuffled symptom, choose exactly

one of the provided shuffled option texts. Base the answer on what the client would endorse, not on what was explicitly disclosed. Return only the required structured response, with no rationale, commentary, paraphrased options, or extra text.

From symptom scores to dialogue. Each PHQ-8 symptom score is translated into private behavioral cues for dialogue generation. The behavior-cue prompt generates short, atomic, naturalistic cues for each PHQ-8 symptom–score cell. It emphasizes lived experience over questionnaire wording and excludes PHQ names, option labels, numeric scores, direct symptom paraphrases, first-person self-report phrasing, and strong cross-symptom contamination. These cues cover affect, functioning, cognition, somatic burden, and interaction style. Candidate cues are first checked within their target cell and then pruned under joint composition tests that sample full PHQ-8 profiles and retain cues whose induced self-reports recover the intended symptom and total scores. After pruning, each PHQ-8 symptom–score cell contains ten validated behavioral cues. At generation time, each symptom contributes five cues. Fractional scores are represented by sampling from adjacent integer severity levels in proportion to the fractional value. The client simulator receives these cues as private self-beliefs and is explicitly instructed not to reveal questionnaire language or numeric labels.

The conversation generator conditions on the private symptom cues, the PsychEval-derived client profile, memory from earlier visits, and turn-level constraints calibrated from RealCBT [35]. Counselor prompts receive only public dialogue context, a short carryover note, retrieved memories, the session focus, the current counseling-move target, and a RealCBT-derived turn-form card. Client prompts receive the public context plus private symptom cues, stable background, session beat guidance, carryover, retrieved memories, and the same style of turn-form card. RealCBT provides surface calibration. Speaker- and stage-specific statistics determine sentence-count patterns, turn length, question form, discourse starters, and lexical overlap, while the client profile and symptom cues determine the clinical content. These constraints embed symptom cues in counseling exchanges.

In the main set, visible counselor and client turns are generated with Qwen3.5-35B-A3B under moderate sampling, using temperature 0.6, nucleus mass 0.95, top-k sampling at 20, and a 512-token cap per turn. Longitudinal run assignment uses seed 42, and the inference backend uses a fixed server seed. Each run contains five sessions with 20 counselor turns and 20 client turns per session. Generated turns must be in English, role-consistent, non-empty, non-repetitive, and free of questionnaire or score leakage and premature closure. Violations trigger targeted repair attempts.

Latent target and self-report supervision. In COUNSELARC, the latent PHQ-8 state is known by construction and serves as the controlled evaluation target. Models train on simulated self-report scores and are evaluated against the latent state that drives dialogue generation. The latent values serve as private generation conditions, while the visible dialogue conveys symptom evidence through client behavior and language.

After each completed session, the client simulator privately completes PHQ-8 using the same behavior cues, stable profile, carryover note, and transcript. Symptom and option order are deterministically shuffled for each pass, and the prompt requires exact option texts without rationale or extra prose. We generate five deterministic self-report passes and average them symptomwise. Invalid structure, missing symptoms, or paraphrased options trigger a repair prompt. Sessions without a valid self-report are excluded from split construction. Session summaries and self-report responses use deterministic decoding, with 1024 tokens allowed for summaries and 1024–1536 tokens for self-report responses depending on the generation setting.

Dataset statistics and use. The complete main Qwen-generated set used in all reported experiments contains 3,690 five-session trajectories and 18,450 labeled counseling sessions. A trajectory-level audit assigns generated clients to fluctuating, stable, improving, and worsening patterns, with all four classes represented. All model experiments use trajectory-level partitions, keeping sessions from each trajectory in one partition.

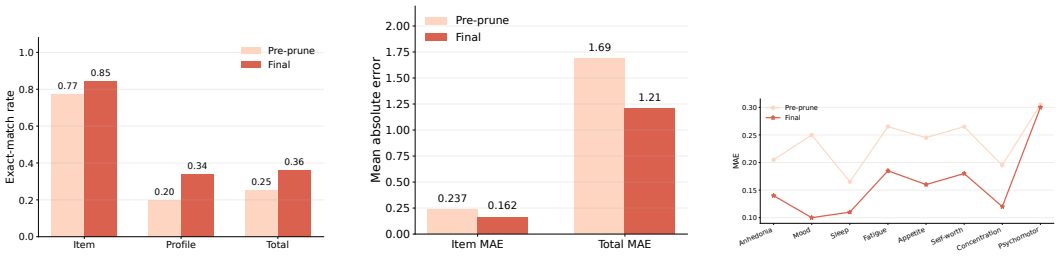
Filtering precedes split construction. Trajectory reconstruction keeps strict five-visit PSYCHE-D participants, NHANES decomposition uses only complete valid symptom-response rows, the cue bank is pruned to a uniform ten cues per symptom–score cell, and generated runs must contain a valid transcript, latent target, and simulated self-report. The main-set split manifest is generated over the resulting 3,690 usable trajectories, assigning 2,583 to training, 369 to validation, and 738 to testing. Split assignment is produced by sorting run ids, applying a seed-42 random shuffle within the

COUNSELARC parent group, and assigning the 0.7/0.1/0.2 train/validation/test ratios at the trajectory level. The manifest records the split for each run id.

COUNSELARC is used to evaluate current-session PHQ-8 prediction under prescribed symptom trajectories, partial disclosure, and cross-session continuity. We therefore pair experiments on COUNSELARC with evaluation on DAIC-WOZ [8], which provides a real single-session anchor for text-based depression assessment.

C.2 Self-report fidelity audit

Pruning the candidate behavior-cue bank substantially improves agreement between simulated self-reports and prescribed latent PHQ-8 states across all granularity levels. Figure 6 compares pre-pruning and final cue banks under the runtime composition protocol (five sampled cues per symptom), reporting exact-match rates, aggregate item and total-score MAE, and per-symptom MAE against the hidden PHQ-8 vector. Pruning raises exact-match rates by 8 percentage points at the item level, 14 points at the full-profile level, and 11 points at the total-score level, while reducing item MAE by 31.6%, total-score MAE by 28.4%, and error across every individual PHQ-8 symptom. Section 5.2 and Figure 2a summarize this fidelity audit.



(a) Item, profile, and total-score exact-match rates.

(b) Item and total-score MAE.

(c) MAE for each PHQ-8 symptom.

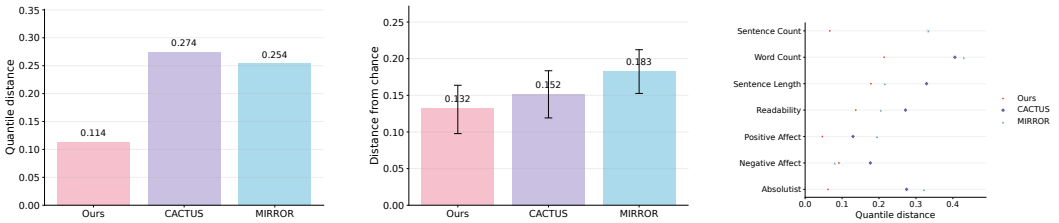
Figure 6: Self-report fidelity before and after cue pruning, with five cues per symptom.

C.3 RealCBT similarity audit

COUNSELARC matches the surface and affective characteristics of real counseling dialogue in RealCBT more closely than existing synthetic counseling corpora [11, 16, 35]. Figure 7 evaluates alignment with RealCBT for COUNSELARC, CACTUS, and MIRROR using aggregate psycholinguistic quantile distance, pointwise judge distance from chance (accounting for systematic judgment inversion), and feature-level quantile distances over speaker-specific turn length, sentence structure, readability, affective language, and absolutist language. COUNSELARC reduces aggregate psycholinguistic distance by 58.4% relative to CACTUS and 55.1% relative to MIRROR, lowers judge distance from chance by 13.2% and 27.9%, respectively, and achieves the closest RealCBT alignment on six of the seven psycholinguistic features (with MIRROR slightly closer only on negative-affect ratio). Section 5.2 and Figures 2b–2c highlight these comparisons in the main paper.

D Additional experiment settings

Hyperparameters. Table 7 reports the reference configuration selected on the main COUNSELARC validation split. The model in this configuration has approximately 0.972M trainable parameters. On LongCounsel-8-GPT54-Mini, memory uses $S{=}16$ slots, history dropout is $p_{\text{hist}}{=}0.2$, and $\lambda_{\text{sym}}{=}0$. On DAIC-WOZ, the encoder has three layers, $\lambda_{\text{sym}}{=}0$, and memory is bypassed because no prior session is available. The retained-client-turn cap is $N_{\max}{=}80$ on DAIC-WOZ. Every longitudinal session contains 20 client turns and fits within this cap. The frozen E-DAIC transfer evaluation retains the DAIC-WOZ-trained model and its selected settings. All default comparisons disable the false-negative penalty ($\lambda_{\text{FN}}{=}0$). Appendix E.10 reports the separate optional-penalty experiment.



(a) Aggregate psycholinguistic quantile distance.

(b) Judge distance from chance.

(c) Feature-level quantile distance.

Figure 7: Dialogue similarity to RealCBT for COUNSELARC, CACTUS, and MIRROR. Lower values indicate closer alignment under each measure.

Table 7: EmoTrack reference hyperparameters for the main COUNSELARC set, with the DAIC-WOZ client-turn cap also listed.

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 client turns $N_{\max}$	80	Symptom-loss weight λ_{sym}	0.5
PHQ-8 items J	8	False-negative weight λ_{FN}	0

Training. We train with AdamW [19] (lr $= 10^{-3}$, 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 on validation loss. Configurations are selected by validation MAE as specified in Section 6.1. The memory configuration uses the immediately preceding session, with longer scopes evaluated separately in Appendix E.7. Training takes approximately 5 minutes per seed on a single GPU.

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. Our baseline reproductions match or improve upon published clinical-based results on DAIC-WOZ, while training-based baselines exhibit higher error and seed sensitivity under the official split. Table 2 reports the five-seed test MAE on DAIC-WOZ for all reproduced baselines using released algorithms or author code. For clinical-based baselines, our five-seed mean is 3.0% below the 2.91 MAE reported for AIDA [15] and 12.5% below the 3.46 MAE reported for LMIQ [31].

For training-based baselines, checkpoint and initialization sensitivity on the small DAIC-WOZ split leads to higher multi-seed error than single-run published scores. Relative to the 3.80 test MAE reported by Lau et al. [14], our released dual-encoder run with all-mpnet-base-v2 has a five-seed mean 16.8% higher—spanning from 4.4% above on its best seed to 42.6% above on a seed that selects a near-constant epoch-0 checkpoint—while the Qwen3-Embedding variant averages 7.7% above the published score with a best seed 4.4% below it. Relative to the 3.78±0.13 test MAE reported by Milintsevich et al. [22], our five-seed mean is 17.4% higher (with the best seed 13.6% higher and a direct author-repository audit 19.5% higher), reflecting an author pipeline that excludes one training participant with missing item labels and two corrupted development transcripts before selecting checkpoints by development loss.

Adapting EnsemBERT [29] from social-media BDI-II prediction to clinical PHQ-8 estimation requires hyperparameter tuning to prevent constant-output collapse. We evaluate test MAE on DAIC-WOZ

before and after a 729-configuration grid search (six optimization and capacity hyperparameters over three values each across five seeds) on the official training and validation splits, keeping the architecture, frozen representation model, retrieval model, and loss weights fixed. Whereas two of the five all-mpnet-base-v2 seeds collapse to all-zero predictions under the published settings, the selected configuration (learning rate 10^{-3}, weight decay 10^{-4}, 30 retrieved posts, 256 dense units, two attention heads, and batch size 64) eliminates constant-output seeds and reduces five-seed test MAE by 27.4% in Table 2, and is applied unchanged across all datasets and embedding variants.

E Additional experiments

This appendix reports the complete cross-dataset method comparison, extended component and hyperparameter ablations, clinical screening and per-symptom bias analyses, longitudinal transition and trajectory-type stratifications, and inference runtime benchmarks. Unless otherwise noted, results are the five-seed mean ± sample standard deviation of test MAE across seeds $0, \ldots, 4$, where † marks the configuration selected by validation MAE and bold denotes the best test mean within each column.

E.1 Complete method comparison

EmoTrack delivers the most consistent accuracy across single-session, transfer, and longitudinal regimes, bridging the gap between clinical-based and training-based baselines. Table 8 reports the five-seed total-score test MAE across all five evaluation sets. Whereas clinical-based methods lead the baselines on real interviews and training-based methods lead on synthetic corpora, EmoTrack ranks first on DAIC-WOZ (12.5% lower MAE than AIDA) and the complementary set (1.9% lower than EnsemBERT with Qwen3-Emb), ranks second under E-DAIC transfer (6.5% behind LMIQ and at least 24.6% below all trained baselines), and remains within 1.1% and 5.2% of the top trained predictors on COUNSELARC and LongCounsel-8-Luna. Section 6.2 and Figure 3a summarize the aggregate comparison.

Table 8: Total-score MAE across all evaluation sets, with the lowest mean in each column in bold. Values are mean ± sample standard deviation over five seeds.

Method	**Overall MAE ↓**				
	COUNSELARC	LC8-LUNA	GPT-5.4-MINI COMP.	DAIC-WOZ	DAIC-WOZ → E-DAIC
AIDA (now)	3.2210 ± 0.0068	3.4932 ± 0.0045	4.0689 ± 0.0133	2.8234 ± 0.2099	4.2538 ± 0.2697
AIDA (all)	3.2517 ± 0.0051	3.5173 ± 0.0083	4.0948 ± 0.0243	–	–
LMIQ (now)	2.9792 ± 0.0117	3.2784 ± 0.0205	4.0095 ± 0.0566	3.0285 ± 0.0460	**3.4722 ± 0.0729**
LMIQ (all)	2.9585 ± 0.0185	3.2285 ± 0.0146	3.9908 ± 0.0513	–	–
Milintsevich et al.	2.6953 ± 0.0211	**2.5588 ± 0.0226**	4.1530 ± 0.0602	4.4370 ± 0.1197	5.0497 ± 0.1233
EnsemBERT	2.9092 ± 0.1128	3.0336 ± 0.0394	4.0038 ± 0.1396	4.1149 ± 0.3339	4.9047 ± 0.3487
EnsemBERT (Qwen3-Emb)	2.7492 ± 0.0343	3.0561 ± 0.3747	3.9184 ± 0.0753	4.8255 ± 0.7277	5.2465 ± 0.5317
Lau et al.	2.6204 ± 0.0194	2.5768 ± 0.0464	4.0689 ± 0.0865	4.4367 ± 0.5744	5.1880 ± 0.3158
Lau et al. (Qwen3-Emb)	**2.5944 ± 0.0232**	2.5985 ± 0.0499	4.0640 ± 0.1034	4.0927 ± 0.2795	5.3641 ± 0.2134
EmoTrack	2.6238 ± 0.0610	2.6913 ± 0.0362	**3.8447 ± 0.1181**	**2.4708 ± 0.3892**	3.6977 ± 0.2665

E.2 Ablation study: encoder–decoder layers

Prediction accuracy exhibits modest sensitivity to encoder–decoder depth across datasets. Table 9 reports total-score test MAE across five transformer depth configurations on COUNSELARC, DAIC-WOZ, and the complementary set. Test means vary by at most 1.9% across depths on COUNSELARC (with the validation-selected 2–4 configuration within 0.6% of the minimum), while the selected 3–4 configuration achieves the lowest error on DAIC-WOZ (up to 6.0% below other depths) and the selected 2–4 configuration remains within 1.2% of the test minimum on the complementary set. Section 4.2 describes the backbone architecture.

Table 9: Effect of encoder–decoder depth on COUNSELARC, DAIC-WOZ, and the complementary GPT-5.4-mini-generated dataset.

Layers		**Overall MAE ↓**		
Encoder	**Decoder**	COUNSELARC	DAIC-WOZ	GPT-5.4-MINI COMP.
1	4	2.6314 ± 0.0371	2.5125 ± 0.2239	$\mathbf{3.8003 \pm 0.0916}$
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	$\mathbf{2.6071 \pm 0.0362}$	2.6053 ± 0.2492	3.8648 ± 0.0460
3	4	2.6163 ± 0.0304	$\mathbf{2.4708 \pm 0.3892}$ †	3.8615 ± 0.1363

E.3 Ablation study: history dropout

Light history dropout improves longitudinal generalization on the full-scale benchmark, whereas aggressive dropout degrades accuracy on both datasets. Table 10 evaluates total-score test MAE across history-dropout rates $p_{\text{hist}} \in \{0.0, 0.1, 0.2, 0.3\}$ on COUNSELARC and the complementary set. On COUNSELARC, the validation-selected rate of 0.1 achieves the lowest error and reduces MAE by 2.2% relative to no dropout; on the complementary set, no dropout yields the lowest test error (1.4% below the validation-selected 0.2), while increasing dropout to 0.3 raises MAE by 1.9% and 4.1% relative to the best setting on each dataset. Figure 5 visualizes these trends in the main paper.

Table 10: Effect of history dropout on COUNSELARC and the complementary GPT-5.4-mini-generated dataset.

History dropout	**Overall MAE ↓**	
	COUNSELARC	GPT-5.4-MINI COMP.
0.0	2.6824 ± 0.0467	$\mathbf{3.7898 \pm 0.1085}$
0.1	$\mathbf{2.6238 \pm 0.0610}$ †	3.8057 ± 0.0622
0.2	2.6434 ± 0.0735	3.8447 ± 0.1181 †
0.3	2.6744 ± 0.0650	3.9462 ± 0.1312

E.4 Ablation study: retained client turns

Bounding the number of retained client turns improves prediction on long clinical interviews by filtering late-session noise without truncating core disclosures. Table 11 reports total-score test MAE on DAIC-WOZ as the client-turn cap $N_{\max}$ varies from 20 to 384. The validation-selected cap of 80 client turns achieves the lowest error, reducing MAE by 4.6% relative to 384 client turns and by 9.3%–10.4% relative to the shorter 40- and 20-turn caps, while still retaining all 20 client turns in every longitudinal session. Figure 5 plots this sweep alongside history dropout.

E.5 Ablation study: memory mechanism

Combining summary and slot-retrieval memory provides the most reliable gains across dialogue generators, even though summary-only memory suffices on the main corpus. Table 12 compares total-score test MAE on COUNSELARC and the complementary set across memory pathway configurations. Relative to no previous memory, the validation-selected combined pathway reduces MAE by 3.2% on COUNSELARC and 0.5% on the complementary set, whereas summary-only memory lowers MAE

Table 11: Effect of the maximum number of retained client turns on DAIC-WOZ.

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

by 3.9% on COUNSELARC (0.8% below the combined pathway) but increases error by 1.9% on the complementary set. Section 4.4 details both pathways.

Table 12: Effect of the memory mechanism on COUNSELARC and the complementary GPT-5.4-mini-generated dataset.

Memory mechanism	Overall MAE ↓	
	COUNSELARC	GPT-5.4-MINI COMP.
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** †

E.6 Ablation study: memory slots

Compact slot capacity achieves the lowest test error on both longitudinal benchmarks, preventing over-parameterization of prior-session context. Table 13 reports total-score test MAE as the number of memory slots S varies from 4 to 16 on COUNSELARC and the complementary set. Four slots yield the lowest test error on both datasets, whereas expanding capacity to 16 slots increases MAE by 2.8% on COUNSELARC and by 1.0% (the validation-selected setting, within a 1.9% overall range) on the complementary set. Section 4.4 defines the slot-compression module.

Table 13: Effect of memory capacity on COUNSELARC and the complementary GPT-5.4-mini-generated dataset.

Memory slots	Overall MAE ↓	
	COUNSELARC	GPT-5.4-MINI COMP.
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 †

E.7 Ablation study: history scope

Extending memory to two prior sessions achieves the lowest error across both longitudinal datasets, outperforming both single-session and unbounded history. Table 14 compares total-score test MAE across historical scopes when feeding the selected prior client-turn embeddings into the unchanged slot compressor. Relative to one prior session, two-session history reduces MAE by 1.4% on COUNSELARC and 2.3% on the complementary set (4.5% and 2.8% below no memory), whereas compressing all prior sessions yields smaller reductions of 0.5% and 1.5%. Section 4.4 describes the default single-session memory configuration.

E.8 Ablation study: symptom-wise loss

Auxiliary symptom-level supervision substantially improves item-level accuracy and slightly lowers total-score error when training data is abundant, whereas aggregate-only supervision is preferred on smaller datasets. Table 15 reports total-score and symptom-level test MAE across symptom-loss weights $\lambda_{\text{sym}} \in \{0, 0.1, 0.3, 0.5, 1.0\}$. On COUNSELARC, the validation-selected $\lambda_{\text{sym}} = 0.5$

Table 14: Effect of history scope on COUNSELARC and the complementary GPT-5.4-mini-generated dataset, with the previous-session compressor unchanged. Bold marks the lowest mean in each column.

History scope	Overall MAE ↓	
	COUNSELARC	GPT-5.4-MINI COMP.
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	$\mathbf{2.5880 \pm 0.0808}$	$\mathbf{3.7559 \pm 0.0977}$
All history	2.6111 ± 0.0742	3.7870 ± 0.1314

reduces symptom MAE by 25.1% and total-score MAE by 0.9% relative to aggregate-only training ($\lambda_{\text{sym}} = 0$), with $\lambda_{\text{sym}} = 1.0$ reducing symptom MAE by a further 1.7%; conversely, $\lambda_{\text{sym}} = 0$ achieves the lowest error on DAIC-WOZ (3.5%–8.3% below nonzero weights) and is also selected by validation on the complementary set (2.6% above the test minimum at $\lambda_{\text{sym}} = 0.3$). Section 4.3 defines the composite loss.

Table 15: Effect of symptom-wise loss on total-score and item MAE. The weight $\lambda_{\text{sym}} = 0$ denotes aggregate-only supervision.

Training objective	Aggregate MAE ↓			Symptom MAE ↓
	COUNSELARC	DAIC-WOZ	GPT-5.4-MINI Comp.	COUNSELARC
$\lambda_{\text{sym}} = 0$	2.6481 ± 0.0370	$\mathbf{2.4708 \pm 0.3892}$ †	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	$\mathbf{3.7454 \pm 0.1019}$	0.4310 ± 0.0125
$\lambda_{\text{sym}} = 0.5$	$\mathbf{2.6238 \pm 0.0610}$ †	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	$\mathbf{0.4178 \pm 0.0131}$

E.9 Clinical-feature extractor

EmoTrack consistently outperforms clinical-based regressors across different upstream LLM feature extractors. We evaluate five-seed test MAE on DAIC-WOZ under Qwen3.5-35B-A3B and GPT-5.6 Luna clinical extraction using a shared reference configuration (two encoder layers, four decoder layers, λ_{sym}=0.5, and a cap of 80 client turns). Under GPT-5.6 Luna extraction, EmoTrack (2.6548 ± 0.1057) reduces MAE by 5.5% relative to matched AIDA (2.8091 ± 0.1003) and by 15.0% relative to LMIQ (3.1223 ± 0.0508); under Qwen extraction, EmoTrack (2.5831 ± 0.2332) reduces MAE by 8.5% relative to AIDA (2.8234 ± 0.2099) and by 14.7% relative to LMIQ (3.0285 ± 0.0460), with EmoTrack's own mean error varying by only 2.8% between extractors. Section 4.1 describes the clinical-feature token representation.

E.10 False-negative penalty

Enabling the optional false-negative penalty in Eq. (4) substantially lowers missed positive screens on the complementary dataset at a moderate cost in false positives and total-score error. Table 16 reports test FNR, FPR, and total-score MAE at the clinical screening cutoff of PHQ-8 ≥ 10 across all baselines and EmoTrack with $\lambda_{\text{FN}} \in \{0, 2.5\}$. Selected by validation FNR over $\lambda_{\text{FN}} \in \{0, 0.5, 1, 2, 2.5, 3, 4, 5\}$, $\lambda_{\text{FN}} = 2.5$ achieves the lowest FNR among all evaluated methods—reducing FNR by 40.4% relative to $\lambda_{\text{FN}} = 0$ and by 28.2% relative to the strongest baseline, Lau et al.—while increasing FPR by 15.6 percentage points and MAE by 5.0%. Section 6.3 and Figure 4 summarize this screening trade-off.

Across the penalty-weight sweep, sensitivity gains saturate at $\lambda_{\text{FN}} = 2.5$ while larger weights degrade score accuracy without further reducing missed positives. Relative to $\lambda_{\text{FN}} = 0$, FNR already drops by 19.8% at $\lambda_{\text{FN}} = 0.5$ before reaching its 40.4% maximum reduction at $\lambda_{\text{FN}} = 2.5$; for $\lambda_{\text{FN}} \in \{3, 4, 5\}$, FNR rebounds by 3.3%–14.7% above the $\lambda_{\text{FN}} = 2.5$ minimum while total-score MAE rises by up to 14.1% above the unpenalized model.

Table 16: Screening comparison on the complementary dataset at a PHQ-8 threshold of 10, with the lowest mean in each column in bold. Values are mean ± sample standard deviation over five seeds.

Method	FNR ↓	FPR ↓	MAE ↓
AIDA (now)	0.7500 ± 0.0000	$\mathbf{0.0778 \pm 0.0037}$	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, $\lambda_{\mathrm{FN}}=0$	0.7080 ± 0.0694	0.1111 ± 0.0372	$\mathbf{3.8447 \pm 0.1181}$
EmoTrack, $\lambda_{\mathrm{FN}}=2.5$	$\mathbf{0.4220 \pm 0.0853}$	0.2667 ± 0.0515	4.0370 ± 0.0691

E.11 Per-item error and directional bias

EmoTrack maintains sub-point accuracy across all eight PHQ-8 symptoms on real clinical interviews, with opposing item-level biases that partially cancel at the total-score level. Table 17 reports per-symptom MAE and mean signed bias (prediction minus reference label) on the 35-participant DAIC-WOZ development split with released item annotations. Every symptom remains at least 24.6% below a one-point MAE—with core affective symptoms (anhedonia and depressed mood) averaging 27.8% lower error than somatic and cognitive symptoms—while negative biases on sleep, fatigue, and appetite offset positive biases on psychomotor change and anhedonia, whose two largest magnitudes match within 0.7%. Section 4.3 describes the item-level prediction head.

Table 17: Per-symptom MAE and signed bias (prediction minus label) of EmoTrack on the 35-participant DAIC-WOZ development split.

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	$+0.4554$

E.12 Current-score error after session-to-session transitions

Historical context improves estimation most when symptom severity persists across visits. We stratify test sessions by the absolute change in latent PHQ-8 total score from the preceding session and compare total-score MAE for EmoTrack with and without one-session memory. When severity is unchanged, memory reduces MAE by 9.2% on COUNSELARC and 5.9% on the complementary set, whereas after shifts of at least five points, memory increases MAE by 2.2% and 3.4%, respectively. Section 4.4 and Appendix E.13 connect this transition behavior to memory design and trajectory-level performance.

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% [44]. In a separate year-long study, initial and intermediate PHQ-9 states were the most important predictors of later deterioration [20].

E.13 Current-score error by trajectory type

EmoTrack achieves the lowest error on stable depression trajectories and remains closely competitive on improving trajectories, while rapid symptom shifts favor memory-free encoders. Table 18 stratifies five-seed test MAE on COUNSELARC across improving, worsening, stable, and fluctuating PHQ-8 trajectory types. On stable trajectories—the largest clinical subgroup—EmoTrack ranks first and reduces MAE by 5.8% relative to the strongest baseline, Lau et al. (Qwen3-Emb), while trailing that baseline by only 1.3% on improving trajectories (7.0% on fluctuating and 10.3% on worsening trajectories) and outperforming all clinical-based methods across every group. Appendix E.12 examines the corresponding session-to-session transition effects.

Table 18: Test MAE on COUNSELARC by PHQ-8 trajectory type. Improving and worsening denote endpoint changes of ≤ -5 and ≥ 5 points, fluctuating denotes a range ≥ 5 with absolute endpoint change < 5, and stable denotes a range < 5.

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

E.14 Runtime and inference cost

With cached clinical features and client-turn embeddings, EmoTrack completes downstream test-set inference in under two minutes per benchmark, avoiding the multi-hour cost of repeated transcript encoding. Table 19 compares single-GPU inference wall-clock hours on COUNSELARC and DAIC-WOZ across all evaluated methods (excluding upstream feature extraction and training). Whereas hierarchical and dual-encoder baselines [14, 22] repeatedly encode full transcripts, EmoTrack reduces inference time by 99.5%–99.9% on COUNSELARC and 98.4%–99.4% on DAIC-WOZ while remaining within seconds of lightweight tabular regressors. Section 6.1 and Appendix D detail the hardware and evaluation protocol.

Table 19: Inference time in single-GPU hours, using cached inputs for EmoTrack. Values are mean ± standard deviation.

Method	**Inference**	
	COUNSELARC	**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

F Limitations

Real longitudinal clinical validation. Real multi-session counseling data with standardized session-level PHQ supervision are needed to validate longitudinal performance and the memory module in

clinical settings. Our longitudinal evaluation uses synthetic counseling, while DAIC-WOZ provides real single-session interviews. We plan to extend the evaluation with clinical partners. The synthetic self-reports remain imperfect approximations of their generating states, and the dialogue audits measure selected aspects of similarity to RealCBT. The transition and trajectory analyses evaluate current-score error, leaving change-direction accuracy and complete-trajectory recovery untested.

Participant-level uncertainty. The observed DAIC-WOZ improvement has substantial participant-level uncertainty. We average each participant's prediction over five seeds, compare absolute errors against AIDA, and bootstrap participants for a 95% confidence interval. EmoTrack reduces MAE by 0.3324 and improves absolute error for 28 of 47 participants (59.6%), but the reduction interval is $[-0.1581,\ 0.8154]$ and a paired permutation test gives $p = 0.1901$. This estimate scores seed-averaged predictions, while Table 8 averages separately scored runs, yielding a mean difference of 0.3526. The interval spanning zero leaves the population-level improvement uncertain.

History after severity changes. Previous-session memory can retain outdated evidence after changes of at least five PHQ points (Appendix E.12). The current gates do not fully suppress this effect. Future memory mechanisms should account explicitly for changes between sessions, with evaluation on real longitudinal data.

Upstream model and implementation dependence. Broader evaluation is needed to establish sensitivity to upstream model choice. The default configuration uses open-source embedding and clinical-feature models. Appendix E.9 also evaluates GPT-5.6 Luna as a proprietary extractor, under which EmoTrack retains a mean-MAE advantage over AIDA. This comparison covers one proprietary extractor and a separate reference configuration on one dataset. The cached-input timing in Table 19 excludes extraction and training, leaving end-to-end compute cost unmeasured. Baseline comparisons also depend on preprocessing, checkpoint selection, and evaluation protocols, as documented in Appendix D.

Item validity and clinical use. Item outputs have not been independently validated as symptom measurements, particularly under aggregate-only supervision. The clinical features are LLM-derived representations, and the PHQ-8 queries impose an architectural prior on total-score prediction. The public DAIC-WOZ test split lacks item labels, and the development-set audit finds opposing item biases that can cancel in the total score (Appendix E.11). Symptom-level use requires item calibration and validation. The screening analysis in Section 6.3 and Appendix E.10 already evaluates the optional false-negative penalty, which increases false positives and score error as it reduces missed positives. Clinical use requires real-patient evaluation with application-specific operating costs, calibration, uncertainty estimation, privacy safeguards, and human review. Diagnostic and intervention uses remain unvalidated.

G Asset licenses and data statement

The EmoTrack source code is publicly available at github.com/Xtra-Computing/EmoTrack under the Apache License 2.0. The evaluation data comprise COUNSELARC and two datasets from LongCounsel-8 [17], LongCounsel-8-Luna and LongCounsel-8-GPT54-Mini. They are available at huggingface.co/datasets/hiddensev/LongCounsel-8 under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. This license preserves the noncommercial condition stated by PSYCHE-D and PsychEval. Table 20 gives a compact summary of new and existing asset terms. We cite all original assets and do not redistribute restricted third-party datasets.

Access conditions and safeguards. The dataset card provides citation instructions, edition descriptions, split manifests, and access instructions. Users request access on Hugging Face and authenticate after approval to download these datasets. The code release includes license files and reproduction instructions. The dataset is intended for research. Users must respect upstream access restrictions and avoid claims about real individuals based on synthetic records.

Table 20: Asset licenses, access terms, and intended use.

Asset	Use	License/access terms
EmoTrack code	Training/evaluation implementation, baseline wrappers, and re-implementations.	Release repository: https://github.com/Xtra-Computing/EmoTrack; Apache License 2.0.
COUNSELARC	Synthetic benchmark for transcript-based PHQ-8 prediction and longitudinal tracking.	CC BY-NC 4.0. Research use only; not for clinical diagnosis, autonomous triage, or claims about real individuals.
DAIC-WOZ [8]	Real single-session PHQ-8 benchmark.	USC ICT application-based access for academic/non-profit research; used under provider terms and not redistributed. `https://dcapswoz.ict.usc.edu/`
NHANES DPQ_L [5]	Symptom co-occurrence source for COUNSELARC construction.	CDC/NCHS public-use data under the NCHS Data User Agreement; no re-identification or identifiable-data linkage. `https://www.cdc.gov/nchs/policy/data-user-agreement.html`
PSYCHE-D [20]	Aggregate depression-change trajectory source.	CC BY-NC 4.0. Used for trajectory reconstruction; raw source files are not redistributed. `https://zenodo.org/records/5085146`
PsychEval [24]	Case/background and session-stage structures.	CC BY-NC 4.0. The source is cited and is not redistributed as a separate raw asset. `https://github.com/ECNU-ICALK/PsychEval/blob/main/LICENSE`
RealCBT [35]	Dialogue-form calibration and similarity audit.	Provider/repository terms; cited and not redistributed as raw sessions. `https://gitlab.com/xiaoyi.wang/realcbt-dataset`
Qwen models [28]	Qwen3-Embedding-8B for embeddings and Qwen3.5-35B-A3B for feature extraction.	Apache License 2.0 as listed on the model cards. `https://huggingface.co/Qwen/Qwen3-Embedding-8B`; `https://huggingface.co/Qwen/Qwen3.5-35B-A3B`
Baseline methods [14, 15, 22, 29, 31]	Reimplemented or adapted comparisons.	Original papers are cited; our released wrappers/reimplementations are under Apache License 2.0, and external author repositories are referenced but not redistributed.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and Section 1 state the method, evaluation settings, and main MAE claim, which are supported by Sections 4 and 6. Appendix C documents the longitudinal dataset. The scope is bounded by the limitations discussed in Appendix F.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Appendix F discusses the need for real longitudinal clinical validation, the use of open-source upstream models, and the distinction between PHQ-8 prediction and deployable clinical decision support.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper does not present theoretical results or proofs. The equations in Section 4 define the model and training objective.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Sections 4 and 6 and Appendices B, D, E, and G provide the model, splits, metrics, baselines, hyperparameters, asset terms, and release plan. Section 6.1 links the evaluation data, and Appendix G describes data access and the planned code release.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: Appendix G documents the code and data release under Apache 2.0 and CC BY-NC 4.0, respectively, and provides the GitHub code URL and dataset URL together with access conditions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Section 6.1 specifies datasets, splits, upstream models, feature extraction, evaluation protocol, and baselines. Appendix D provides hyperparameters and training details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: Section 6.1 states that results are reported as five-seed means with standard deviations, and the result tables report mean $\pm$ sample standard deviation. The seed variability is described through matched-seed feature extraction and downstream training. Appendix F adds a participant-level bootstrap interval and paired permutation test for the main DAIC-WOZ comparison.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Appendix D reports the local server system, GPU, CPU, and memory. Appendix E.14 and Table 19 provide detailed per-method inference time and single-GPU-hour estimates for the main comparison runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: Appendix F discusses clinical validation, privacy safeguards, and human-in-the-loop review. Appendix G documents asset terms, intended use, and release conditions.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes].

Justification: Section 1 motivates potential use for monitoring depression severity and supporting timely human review. Appendix F discusses risks around clinical validation, privacy safeguards, false negatives, and human-in-the-loop deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes].

Justification: Appendix G specifies release licenses, intended use, access conditions, no redistribution of restricted upstream assets, and safeguards against clinical misuse of the synthetic mental-health benchmark. Appendix F further discusses deployment safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: Existing datasets, models, and baselines are cited throughout Sections 2 and 6 and Appendix C. Appendix G summarizes the applicable licenses or provider access terms for DAIC-WOZ, NHANES/DPQ, PSYCHE-D, RealCBT, PsychEval-derived assets, Qwen models, and baseline implementations.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes].

Justification: COUNSELARC is documented in Section 6.1 and Appendix C. Appendix G adds a data statement covering license, intended use, limitations, access conditions, and release safeguards.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not report new crowdsourcing experiments or new human-subject data collection. It uses synthetic data and existing datasets, whose asset terms should be documented under the licenses and new-assets checklist items.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not conduct new human-subject or crowdsourcing studies. Existing human-derived corpora are used as external assets, and their access conditions and original approvals should be covered through the asset documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: Section 4 describes LLM-extracted clinical features as part of EmoTrack, and Section 6.1 identifies the Qwen models used for feature extraction and embeddings. Appendix C describes LLM-based synthetic dialogue construction.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.