Probabilistic Textual Time Series Depression Detection

Anonymous ACL submission

Abstract

Accurate and interpretable predictions of depression severity are essential for clinical decision support, yet existing models often lack uncertainty estimates and temporal modeling. We propose PTTSD, a *Probabilistic Textual Time Series Depression Detection* framework that predicts PHQ-8 scores from utterance-level clinical interviews while modeling uncertainty over time. PTTSD includes sequence-to-sequence and sequence-to-one variants, both combining bidirectional LSTMs, self-attention, and residual connections with Gaussian or Student's-t output heads trained via negative log-likelihood. Evaluated on E-DAIC and DAIC-WOZ, PTTSD achieves state-of-the-art performance among text-only systems (e.g., MAE = 3.85 on E-DAIC, 3.55 on DAIC) and produces well-calibrated prediction intervals. Ablations confirm the value of attention and probabilistic modeling, while comparisons with MentalBERT establish generality. A three-part calibration analysis and qualitative case studies further highlight the interpretability and clinical relevance of uncertainty-aware forecasting.

1 Introduction

Depression remains one of the leading causes of global disability, affecting over 300 million individuals worldwide (WHO, 2017, 2022). Scalable, automated tools for assessing depressive symptom severity offer valuable support in digital therapy and remote care, where access to clinicians is limited. Among these tools, text-based systems that process clinical interviews have shown strong potential for predicting standardized scores such as the PHQ-8 (Kroenke et al., 2009).

Recent methods typically model interview transcripts as sequences of utterances and employ deterministic architectures such as LSTMs, Transformers, or large language models (LLMs) (Mandal et al., 2025; Fang et al., 2023a; Nykoniuk et al., 2025; Sadeghi et al., 2024). These utterance sequences naturally form a *textual time series*, where each utterance provides a temporally ordered observation. However, most existing approaches produce scalar severity estimates without quantifying uncertainty—an important limitation in high-stakes clinical contexts.

We argue that modeling depression severity as a *probabilistic textual time series regression* task enables not only accurate predictions but also interpretable confidence estimates. In this formulation, each utterance contributes to an evolving posterior over severity scores, allowing us to capture *aleatoric uncertainty*—uncertainty arising from inherent noise or ambiguity in the input, such as sparse, contradictory, or ambiguous language. This uncertainty-aware perspective is crucial in clinical natural language processing (NLP), where model confidence can significantly affect downstream decision-making.

We introduce *PTTSD*—a *Probabilistic Textual Time Series Depression Detection* framework that makes temporally grounded, calibrated predictions over PHQ-8 scores from utterance-level sequences. Unlike prior work, PTTSD addresses three key limitations in the field: (1) it produces calibrated uncertainty estimates rather than point predictions; (2) it avoids prompt-based methods, improving reproducibility; and (3) it leverages full transcripts to capture temporal structure. Our model encodes the input with a bidirectional LSTM, self-attention, and residual connections, and produces distributional outputs using Gaussian or Student's-t heads trained via negative log-likelihood loss.

We evaluate PTTSD on the DAIC and E-DAIC benchmarks using high-quality re-transcribed interviews and demonstrate strong performance on standard metrics (e.g., MAE = 3.55, RMSE = 4.77 on DAIC; MAE = 3.85, RMSE = 4.52 on E-DAIC), outperforming recent text-only baselines without relying on prompt engineering or handcrafted fea-

tures. Ablation and sensitivity analyses further validate the contributions of probabilistic loss design, attention mechanisms, and calibration metrics.

Our main contributions are:

- We propose PTTSD, a fully probabilistic sequence model that jointly predicts PHQ-8 scores along with calibrated uncertainty from utterance-level textual time series.
- We train and evaluate PTTSD end-to-end on full interviews without handcrafted prompts, yielding a simple, reproducible modeling pipeline.
- We achieve state-of-the-art results on E-DAIC among text-only models, and provide thorough ablation, calibration, and sensitivity analyses to assess uncertainty quality and model robustness.

2 Related Work

Textual time series modeling has been central to recent efforts in automatic depression detection, especially within clinical interviews and therapy sessions. Prior work has predominantly relied on deterministic neural methods such as LSTMs and attention-based transformers to model temporal dependencies in textual data (Mandal et al., 2025; Fang et al., 2023a; Nykoniuk et al., 2025). These models capture sequential patterns but lack mechanisms to quantify uncertainty over time. While LLMs extract richer textual features (Sadeghi et al., 2024; Chen et al., 2024), most systems remain heuristic or deterministic, focusing on structural or multimodal fusion rather than probabilistic reasoning. In contrast, our fully probabilistic, end-to-end model captures uncertainty directly from raw utterances without handcrafted prompts, emphasizing simplicity and efficiency.

Notably, Qureshi et al. (2019b) use multitask learning with attention mechanisms for joint regression and classification, but do not incorporate uncertainty modeling. Similarly, prompt-based methods such as those of Zhang and Guo (2024) transform depression detection into a few-shot classification task via language model prompting, but still yield single-point predictions. Graph-based architectures (Burdisso et al., 2023; Chen et al., 2024) model discourse-level context across utterances and questions, offering enhanced interpretability and structural awareness, though they too typically omit calibrated uncertainty.

A rare exception is Dia et al. (2024), who propose a stochastic transformer for post-traumatic stress disorder detection, introducing probabilistic components such as stochastic activations to model uncertainty across modalities. However, their work focuses on visual signals and does not address textual time series or PHQ-8 regression. More recently, Zhang et al. (2025) apply a multi-instance learning (MIL) framework to estimate depression severity from long transcripts, assigning confidence scores to depressive cues at the sentence level. While this provides instance-level interpretability, the underlying model is not explicitly probabilistic in the Bayesian sense.

Several recent works have explored fair or calibrated uncertainty estimation. Li and Zhou (2025) propose Fair Uncertainty Quantification (FUQ) for PHQ regression, producing conformal prediction intervals with coverage guarantees across demographic groups. While effective for fairness, FUQ operates at the distributional output level and does not model temporal evolution within interviews. Other systems, such as Mao et al. (2023) and Guo et al. (2022), employ BiLSTMs or Transformers with textual features, sometimes augmented by topic signals, but focus solely on deterministic loss objectives.

3 Probabilistic Textual Time Series Depression Detection

3.1 Problem Formulation

We formalize PHQ-8 prediction as a probabilistic regression task over utterance-level textual sequences. Given a clinical interview transcript consisting of T utterances $\{u_1, u_2, \ldots, u_T\}$, the goal is to predict a scalar depression severity score $y \in \mathbb{R}$ (e.g., the participant's PHQ-8 score), and quantify uncertainty in that prediction.

$$p(y \mid e_{1:T}; \theta)$$

3.2 Data and Preprocessing

We utilize the Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014) and extended DAIC (E-DAIC) (DAIC-WOZ Project, 2019) datasets, which contain anonymized semi-structured interview transcripts and associated PHQ-8 (Kroenke et al., 2009) depression scores. Each participant's data consists of a sequence of utterances extracted from transcript files, along with a PHQ-8 score indicating depression severity. The PHQ-8 (Pa-

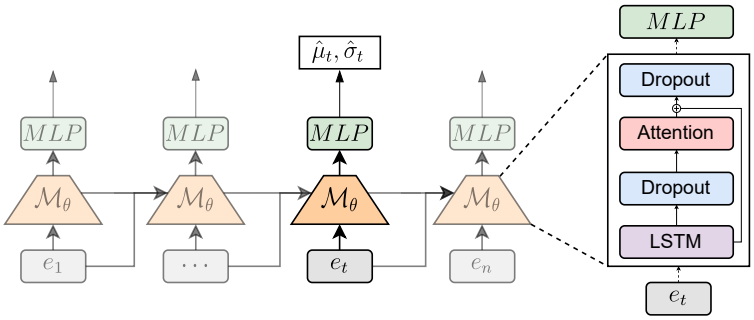


Figure 1: Probabilistic Textual Time Series Depression Detection

tient Health Questionnaire-8) is a standardized self-report instrument with scores ranging from 0 to 24, used to assess depressive symptom severity. More details on the PHQ-8 and DAIC in Appendix A and Appendix B, respectively.

To improve interview transcription fidelity, we reprocessed the original E-DAIC audio using WhisperX (Bain et al., 2023), which provides more accurate word-level alignment and robust speaker diarization compared to the baseline Whisper model (Radford et al., 2023) employed by Sadeghi et al. (2024). We organize utterances into temporal sequences and split the data into training, validation, and test sets using the predefined partitions. During batching, utterances are padded to the batch's maximum length, and an attention mask is constructed to differentiate padded from valid tokens.

3.3 Generating Utterance Embeddings

We represent each utterance using pretrained sentence encoders. We represent each utterance using pretrained sentence encoders. Our primary model uses the `all-MiniLM-L6-v2` Sentence Transformer[1] (Reimers and Gurevych, 2019), a compact model with only 22 million parameters that achieves competitive performance across a wide range of tasks on the Hugging Face MTEB Embedding Leaderboard (Muennighoff et al., 2023). We also evaluate an alternative variant of our model using `MentalBERT` (Ji et al., 2022), a domain-adapted BERT model pretrained on mental health-related corpora[2].

For Sentence Transformers, utterances are encoded directly into fixed-dimensional vectors using mean pooling over token representations. For MentalBERT, we extract the final hidden state of the `[CLS]` token as the utterance-level embedding. In both cases, each utterance u_t is independently mapped to an embedding vector $e_t \in \mathbb{R}^D$. The resulting sequence $(e_1, e_2, \ldots, e_T)$ represents the input utterance series, where T is the number of utterances in the session. These embeddings are stacked into a tensor $\mathbf{X} \in \mathbb{R}^{B\times T\times D}$, where B is the batch size, T the number of utterances, and D the embedding dimension. We propagate attention masks through the entire pipeline to exclude padded positions from contributing to downstream modeling, loss computation, and evaluation.

3.4 Probabilistic LSTM Architecture

PTTSD uses a unified sequence model that supports both sequence-to-sequence (seq2seq) and sequence-to-one (seq2one) prediction modes. Both variants share the same architectural backbone inspired by (Mandal et al., 2025) (Figure 1): a multi-layer bidirectional LSTM followed by a multi-head self-attention layer (Vaswani et al., 2017) with residual connections. Let $\mathbf{X} \in \mathbb{R}^{B\times T\times D}$ be the input utterance embedding sequence. The LSTM encodes this into hidden states $\mathbf{H} \in \mathbb{R}^{B\times T\times H}$, which are then passed through the attention layer to produce refined representations. The final attended representation is obtained via a residual connection:

$$\mathbf{A} = \text{Attention}(\mathbf{H}) + \mathbf{H}$$

Sequence-to-Sequence. In this mode, each time step t yields a predictive distribution $p(y_t \mid e_{\leq t})$. Two feedforward networks (MLPs) map the attended hidden state $\mathbf{a}_t$ to the mean and uncertainty parameters:

$$\hat{\mu}_t = \text{MLP}_\mu(\mathbf{a}_t), \quad \hat{\sigma}_t = \text{softplus}(\text{MLP}_\sigma(\mathbf{a}_t)) + \epsilon$$

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] https://huggingface.co/mental/mental-bert-base-uncased

where $\epsilon > 0$ ensures numerical stability. Depending on the output head, the model may alternatively predict evidential parameters or degrees of freedom for a Student's-t distribution. All predictions are computed in parallel and masked to exclude padded positions.

Sequence-to-One. In this mode, the attended sequence $\mathbf{A}$ is aggregated into a single pooled vector using masked average pooling. This pooled representation is then used to predict a single distribution $p(y \mid e_{1:T})$, matching the session-level annotation granularity of PHQ-8.

3.5 Sequence Modeling and Predictive Distributions

In the sequence-to-sequence variant, the PHQ-8 score is modeled as a time series where the label at time step t is predicted as:

$$p(y_t \mid e_{\leq t}; \theta)$$

where θ denotes the model parameters and $e_{\leq t}$ are the utterance embeddings up to time t. The model is trained non-autoregressively, i.e., without access to past labels $y_{<t}$. In the sequence-to-one variant, a single distribution is predicted for the entire sequence:

$$p(y \mid e_{1:T}; \theta)$$

corresponding to the session-level PHQ-8 target.

We explore two probabilistic output distributions:

Gaussian distribution. The model predicts a mean $\hat{\mu}_t$ and standard deviation $\hat{\sigma}_t$ at each time step, defining the conditional distribution as:

$$p(y_t \mid e_{\leq t}; \theta) = \mathcal{N}(y_t \mid \hat{\mu}_t, \hat{\sigma}_t^2)$$

Student's *t*-distribution. Alternatively, the model may output a location $\hat{\mu}_t$, scale $\hat{\sigma}_t$, and degrees of freedom ν_t, defining:

$$p(y_t \mid e_{\leq t}; \theta) = \text{StudentT}(y_t \mid \hat{\mu}_t, \hat{\sigma}_t, \nu_t)$$

The corresponding probability density function is:

$$f(y \mid \mu, \sigma, \nu) = C(\nu, \sigma) \left[1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}$$

with normalization constant:

$$C(\nu, \sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\,\sigma}$$

3.6 Loss Functions

We train all models using the negative log-likelihood (NLL) of the ground-truth PHQ-8 score under the predicted distribution. For the sequence-to-sequence variant, the loss is computed at each time step and averaged across valid utterances. For the sequence-to-one variant, a single prediction is made per session, and the NLL is computed at the sequence level.

Let $\hat{\mu}_t$ and $\hat{\sigma}_t$ denote the predicted mean and standard deviation at time step t (or $\hat{\mu}, \hat{\sigma}$ in the seq2one case, where $t = 1$). The total sequence loss is:

$$\mathcal{L}_{\text{seq}} = \begin{cases} -\sum_{t=1}^{T} \log p(y_t \mid e_{\leq t}; \theta) & \text{(seq2seq)} \\ -\log p(y \mid e_{1:T}; \theta) & \text{(seq2one)} \end{cases}$$

The batch loss is normalized across participants:

$$\mathcal{L}_{\text{batch}} = \frac{1}{B}\sum_{i=1}^{B} \frac{1}{T_i} \mathcal{L}_{\text{seq}}^{(i)}$$

where T_i is the number of valid utterances for participant i; for seq2one, $T_i = 1$.

For Gaussian outputs, the model predicts a mean and standard deviation, and the corresponding weighted Gaussian NLL loss is:

$$\mathcal{L}_{\text{NLL}} = \sum_{t=1}^{T} \left[\alpha \log(2\pi) + \beta \log(\hat{\sigma}_t^2) + \gamma \cdot \delta_t\right]$$

$$\text{with } \delta_t = \frac{(y_t - \hat{\mu}_t)^2}{\hat{\sigma}_t^2}$$

For seq2one, this reduces to a single-term sum with $t = 1$. The weights α, β, γ control the trade-off between likelihood components and are set to 1 by default.

We optionally use auxiliary or baseline objectives:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu}_t)^2, \mathcal{L}_{\text{MAE}} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{\mu}_t|$$

In the seq2one case, the summation reduces to a single term, as only one label and one prediction exist per sample.

3.7 Training Procedure

PTTSD is trained for 50 epochs using the Adam optimizer with cosine annealing (Loshchilov and Hutter, 2016). The learning rate is initialized at 2e-4 and decays smoothly to a minimum of 1e-4. We batch at the participant level, with each batch containing all utterances from a subset of participants. Early stopping with a patience of 15 epochs is applied based on Dev MAE, and the best-performing model checkpoint is restored. To address label imbalance, we apply a log transformation to the targets during training, with outputs transformed back to the original scale for evaluation.

4 Experiments

4.1 Experimental Setup

Implementation. All models are implemented in PyTorch (Paszke et al., 2019). Padding, batching, and masking ensure that variable-length sequences do not affect loss or metric computations.

Hardware. Training is performed on a single NVIDIA A100-SXM4-80GB GPU with 80GB of GDDR6 VRAM, using CUDA version 12.2.

Runtime. Training PTTSD for 50 epochs on a single NVIDIA A100–80 GB takes ~2h 23min in wall-clock time (≈172 s per epoch). The model has a total 2,703,403 trainable parameters.

Data Splits. We follow the official training, validation, and test splits provided with each dataset. For E-DAIC, the data is partitioned into 163 training, 56 validation, and 56 test participants. For DAIC-WOZ, the official splits include 107 training, 35 validation, and 56 test participants. As described in Section 3.2, all audio is re-transcribed using WhisperX to improve transcription quality and alignment over the original transcripts.

Evaluation Metrics. We evaluate models on both the validation and held-out test sets using mean squared error (MSE) and root mean squared error (RMSE). These metrics quantify average prediction error, with RMSE placing greater emphasis on larger errors due to its squaring operation. This makes RMSE particularly useful for identifying models that minimize not just average error, but also variance in error magnitude. When modeling predictive uncertainty, we additionally report negative log-likelihood (NLL). All metrics are computed over valid (non-padded) utterances only.

Reproducibility. All preprocessing steps, model configurations, and training scripts are made publicly available on GitHub.[3] To account for variability due to random initialization, we report average performance over three runs with different seeds.

4.2 Main Results

Table 1 presents PHQ-8 regression performance on both E-DAIC and DAIC datasets. We compare our PTTSD models across multiple configurations (sequence-to-sequence vs. sequence-to-one; MentalBERT vs. all-MiniLM-L6-v2) against a range of prior text-based approaches.

E-DAIC. Among all text-only systems, PTTSD (sequence-to-sequence with all-MiniLM-L6-v2) achieves the lowest test MAE (3.85) and RMSE (4.52), establishing a new state of the art. Other PTTSD variants, including MentalBERT-based and sequence-to-one configurations, also perform competitively, demonstrating robustness across architecture choices. Earlier works such as Ray et al. (2019) and Rodrigues Makiuchi et al. (2019) attain dev RMSEs of 4.22–4.97, but their test performance is either weaker or unreported. More recent prompt-based models by Sadeghi et al. (2024) use Whisper transcripts and audio-based quality filtering. Their best variant (Pr3+Whisper+AudioQual) reports strong dev MAE (2.85) and RMSE (4.02), but is not text-only due to audio quality gating. Their best text-only test result (Pr3+Whisper) achieves 4.22 MAE and 5.07 RMSE, which PTTSD outperforms by a large margin on both metrics.

DAIC. On the original DAIC dataset, PTTSD again performs competitively, especially in the all-MiniLM-L6-v2 sequence-to-one variant, which achieves the lowest test MAE (3.55) and matches the best test RMSE (4.77) of Fang et al. (2023b). Interestingly, Gong and Poellabauer (2017) reports strong dev performance (MAE 2.77, RMSE 3.54), while test results (MAE 3.96, RMSE 4.99) show a notable drop, which may reflect differences in evaluation protocols or generalization challenges.

4.3 Ablation Studies

Effect of Loss Function. Table 2 compares the impact of different loss functions on validation and test performance. Gaussian NLL yields the best overall balance, achieving low MAE and RMSE across both splits, with particularly strong test

[3] https://github.com/someonedoing-research/PTTSD

Dataset	Method	Dev		Test	
		MAE	RMSE	MAE	RMSE
DAIC	Williamson et al. (2016)	3.34	4.46	–	–
	Gong and Poellabauer (2017)	**2.77**	**3.54**	3.96	4.99
	Yang et al. (2017)	3.52	4.52	–	–
	Stepanov et al. (2018)	–	–	4.88	5.83
	Oureshi et al. (2021)	3.78	–	–	–
	Niu et al. (2021)	3.73	4.80	–	–
	Fang et al. (2023b)	–	–	3.61	**4.76**
	Rohanian et al. (2019)	–	–	4.98	6.05
	Al Hanai et al. (2018)	5.18	6.38	–	–
	Qureshi et al. (2019a)	3.74	4.80	–	–
	PTTSD - *sequence-to-one - MentalBERT*	4.39 (±0.10)	5.47 (±0.43)	3.65 (±0.24)	4.69 (±0.24)
	PTTSD - *sequence-to-sequence - MentalBERT*	4.67 (±0.04)	5.82 (±0.34)	3.92 (±0.54)	4.79 (±0.54)
	PTTSD - *sequence-to-one - all-MiniLM-L6-v2*	3.82 (±0.09)	4.84 (±0.28)	**3.55** (±0.15)	**4.77** (±0.53)
	PTTSD - *sequence-to-sequence - all-MiniLM-L6-v2*	4.59 (±0.07)	5.22 (±0.30)	3.88 (±0.41)	5.10 (±0.92)
E-DAIC	Ray et al. (2019)	–	4.37	4.02	4.73
	Rodrigues Makiuchi et al. (2019) – LSTM	–	4.97	–	6.88
	Rodrigues Makiuchi et al. (2019) – 8 CNN blocks-LSTM	–	4.22	–	–
	Sadeghi et al. (2023)	3.65	5.27	4.26	5.37
	Sadeghi et al. (2024) – Pr3+Whisper	3.17	4.51	4.22	5.07
	Sadeghi et al. (2024) – Pr3+Whisper+AudioQual	**2.85**	**4.02**	3.86	4.66
	PTTSD - *sequence-to-one - MentalBERT*	3.56 (±0.01)	4.45 (±0.07)	4.18 (±0.05)	5.23 (±0.13)
	PTTSD - *sequence-to-sequence - MentalBERT*	3.55 (±0.14)	4.58 (±0.20)	4.20 (±0.03)	5.39 (±0.08)
	PTTSD - *sequence-to-one - all-MiniLM-L6-v2*	3.60 (±0.13)	4.76 (±0.14)	4.58 (±0.50)	5.87 (±0.92)
	PTTSD - *sequence-to-sequence - all-MiniLM-L6-v2*	3.47 (±0.02)	4.57 (±0.04)	**3.85** (±0.04)	**4.52** (±0.38)

Table 1: Evaluation of PHQ-8 regression performance across text-only models on the DAIC and E-DAIC datasets. Bold indicates best performance within each dataset.

Loss	Dev		Test	
	MAE	RMSE	MAE	RMSE
Gaussian NLL	3.4440	4.5293	3.8603	5.0219
Student-t NLL	3.6637	4.9328	3.9294	5.1488
MAE	3.6427	4.8091	4.1885	5.4407
MSE	3.6398	4.9845	3.6694	4.8760

Table 2: Loss function comparison on dev/test sets (E-DAIC, single run).

MAE (3.86). Student's-t NLL performs comparably but with slightly worse calibration and higher RMSE, likely due to the added complexity of estimating the degrees of freedom.

MAE and MSE losses exhibit inconsistent behavior: while MSE achieves the lowest test MAE (3.67), it performs worse on the dev set and yields the highest test RMSE among all probabilistic losses. The MAE loss underperforms across all metrics, suggesting it is less effective for learning stable sequence-level representations in this setting.

These results highlight that Gaussian NLL offers the most reliable and generalizable performance when modeling uncertainty in PHQ-8 prediction from textual time series.

Effect of the Model Architecture. We conduct an ablation study to assess the contribution of individual architectural components in our probabilistic LSTM sequence-to-sequence model. Each ablation variant disables a specific component—attention, residual connections, or the variance prediction head—while all other settings are held constant. Models are trained for 20 epochs (rather than the full 50 used in main experiments) to accelerate comparison. Evaluation is performed on the test set using mean absolute error (MAE) and root mean squared error (RMSE). Full experimental details are included in Appendix C.

Table 3 and Figure 2 illustrate the effects of disabling different components. Removing self-attention yields the largest degradation in perfor-

Variant	MAE	Δ MAE (%)	RMSE	Δ RMSE (%)
Full Model	6.32	–	8.10	–
- w/o Attention	7.74	+22.48	9.74	+20.24
- w/o Residual	7.19	+13.78	8.96	+10.53
- w/o Variance Head	5.98	−5.37	7.21	−10.99

Table 3: Ablation of architectural components (Gaussian NLL on test set). Absolute scores and percentage change relative to the full model.

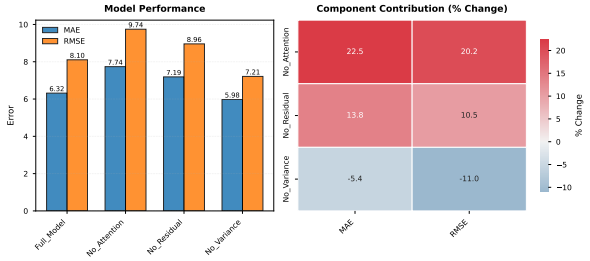


Figure 2: Ablation results

mance, increasing MAE by 22.5% and RMSE by 20.2%, confirming its importance for modeling long-range dependencies across utterances.

Omitting residual connections also leads to noticeable performance drops (MAE +13.8%, RMSE +10.5%), suggesting that residual pathways contribute to stable training and effective information flow across layers.

Interestingly, removing the variance prediction head simplifies the model and yields slightly better raw error metrics (MAE –5.4% and RMSE –11.0%), likely because the model reverts to a simpler deterministic objective. However, this comes at the cost of losing uncertainty estimation—a core benefit in clinical decision support.

Overall, the full model offers the best trade-off between predictive accuracy and uncertainty modeling, with ablations confirming the value of self-attention, residuals, and probabilistic output heads.

4.4 Hyperparameter Sensitivity

α	β	γ	NLL (Dev)	NLL (Test)	Comments
1	1	1	1.3129	1.1934	standard NLL
1	2	1	1.7854	1.4865	uncertainty-averse
1	1	2	1.2777	1.3189	error-focused
1	1	0.5	2.2163	2.0316	calibration-first

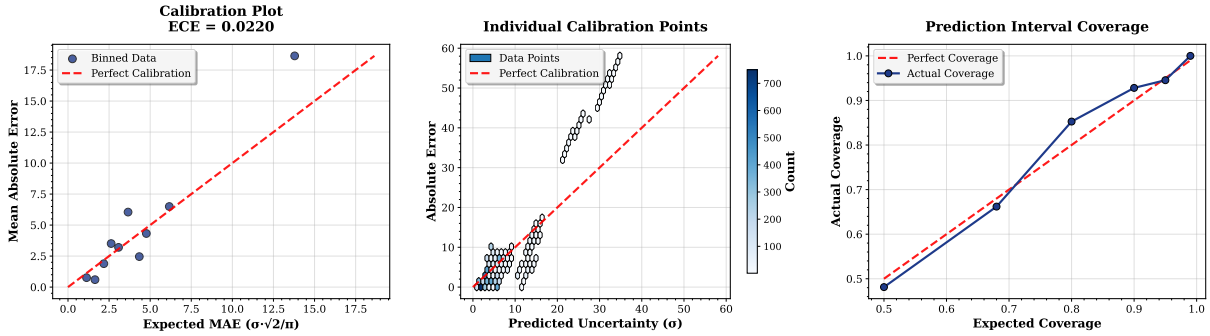
Table 4: Sensitivity analysis of Gaussian NLL loss weighting parameters α, β, and γ.

Table 4 presents the effect of varying the NLL weighting parameters β (log-variance term) and γ (normalized squared error term), with α held constant as it weights the constant term in the NLL and hence does not influence the model's gradients or learning dynamics. The standard setting ($\beta = \gamma = 1$) yields the best overall performance on the test set (NLL = 1.1934), indicating a balanced trade-off between data fit and uncertainty modeling. Increasing β to 2 ("uncertainty-averse") substantially increases NLL on both development and test sets, suggesting that heavily penalizing predicted variance harms calibration and leads to underconfident predictions. Conversely, increasing γ to 2 ("error-focused") improves the development NLL slightly but increases test NLL, indicating overfitting to the training signal. Reducing γ to 0.5 ("calibration-first") degrades both development and test NLLs, likely due to underemphasis on prediction accuracy. The results suggest that aggressive reweighting of either term destabilizes the trade-off between sharpness and calibration, and that the default Gaussian NLL ($\beta = \gamma = 1$) remains the most reliable setting across validation and test sets.

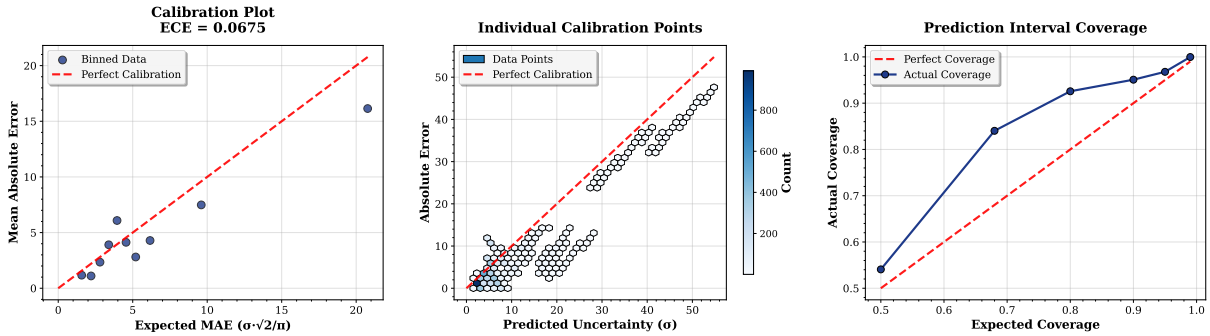
4.5 Uncertainty Calibration Analysis

Accurate uncertainty quantification is critical in clinical NLP, where predictions may inform sensitive decisions. We evaluate the calibration of PTTSD using Expected Calibration Error (ECE), empirical coverage, and visual diagnostics (Figure 3), comparing models trained with Gaussian NLL and MSE losses to understand how probabilistic modeling impacts calibration quality. The figure includes three subplots: (1) a binned calibration plot comparing mean predicted standard deviation (x-axis) and mean absolute error (y-axis) across uncertainty bins, with deviations from the diagonal summarized by the Expected Calibration Error (ECE); (2) a scatter plot of individual predictions showing predicted uncertainty versus observed error; and (3) a coverage plot showing the proportion of ground truth values falling within model-predicted confidence intervals. Perfect calibration aligns with the red diagonal in each plot. We compute the ECE as the average absolute deviation between nominal confidence levels and the actual coverage rates observed at those levels. Specifically, we compare the proportion of ground truth values falling within the model's prediction intervals (e.g., 68%) to the expected theoretical value.

Figure 3 compares calibration results for PTTSD trained with Gaussian NLL and MSE. The Gaussian NLL model achieves a low ECE of 0.0220 and closely approximates ideal 68% coverage (66.2%), indicating strong and informative calibra-



(a) Probabilistic Gaussian NLL ($\alpha = \beta = \gamma = 1$)



(b) Deterministic MSE

Figure 3: Calibration analysis of PTTSD seq2seq on DAIC test set (Gaussian NLL vs. MSE)

tion. While slightly overconfident, it adapts uncertainty to input ambiguity, producing sharp yet reliable intervals. In contrast, the MSE-trained model shows poorer average calibration (ECE = 0.0675) and is significantly underconfident, with coverage far exceeding the nominal threshold (84.0%), resulting in overly wide and less informative intervals. Overall, the Gaussian NLL model yields better-aligned uncertainty estimates balancing sharpness and reliability—critical for clinical NLP applications where actionable confidence matters.

To illustrate the interpretability of our uncertainty estimates, Appendix D presents case studies visualizing predicted PHQ-8 intervals over time, showing how the model adjusts confidence based on input ambiguity and severity dynamics. We also observe a strong correlation between predicted uncertainty and absolute error (Pearson $r = 0.88$, Spearman $\rho = 0.64$; Appendix E.2), confirming that uncertainty estimates reflect prediction quality.

Combined, the calibration metrics and case studies demonstrate that PTTSD produces informative and actionable uncertainty. Unlike point-estimate models, PTTSD adapts its confidence to the input, indicating when predictions are reliable and when caution is warranted. This is especially critical in clinical NLP, where decisions depend not just on *what* is predicted, but also on *how sure* the model is. By providing sequence-level uncertainty that evolves with dialogue, PTTSD supports transparent and interpretable assessment for real-world mental health screening and triage.

5 Conclusion

We introduced PTTSD, a probabilistic neural framework for predicting PHQ-8 depression severity from utterance-level text. PTTSD models calibrated uncertainty using Gaussian and Student's-t distributions and integrates bidirectional LSTMs, self-attention, and residual connections. It requires no handcrafted features or prompts, making it suitable for clinical deployment. Experiments on DAIC and E-DAIC show that PTTSD achieves state-of-the-art performance among fully automatic text-only systems, while providing reliable uncertainty estimates. Ablations and calibration analyses confirm the contributions of attention, probabilistic output heads, and balanced loss weighting. These findings support the utility of uncertainty-aware textual time series modeling in clinical NLP. Future work will explore multimodal extensions and clinical validation.

Limitations

While PTTSD offers promising results in predictive accuracy and uncertainty modeling, several limitations remain. First, the framework relies solely on textual data. Although effective, it does not leverage multimodal cues such as vocal prosody or facial expressions, which are known to be informative for assessing mental health. Second, the E-DAIC dataset contains fewer than 300 participants, and further reduction due to filtering and partitioning limits the statistical power and generalizability of our findings to broader clinical settings. Third, the interviews in E-DAIC are conducted with a virtual interviewer ("Ellie") operated in a Wizard-of-Oz setup rather than a real clinician, which may affect the ecological validity of the speech data and limit applicability to authentic client–clinician interactions. In terms of modeling, we encode utterances independently using pretrained language models without context-aware finetuning, potentially overlooking local coherence or discourse-level cues. Furthermore, while PTTSD provides distributional predictions, we do not assess its clinical utility or decision-support value; human-centered evaluations with therapists or end users are needed to determine the interpretability and trustworthiness of predicted uncertainty. Finally, although we evaluate calibration quantitatively, we do not study how uncertainty scores might be perceived or utilized by clinicians in real-world settings. Future work should address these limitations by incorporating multimodal signals, validating on therapist–client dialogues, and evaluating the human trust and usability of uncertainty-aware predictions.

References

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

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. In *INTERSPEECH 2023*, pages 3617–3621.

Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. 2024. Depression detection in clinical interviews with LLM-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8181–8194, Mexico City, Mexico. Association for Computational Linguistics.

DAIC-WOZ Project. 2019. Extended distress analysis interview corpus–wizard of oz (e-daic). Extended DAIC Database, downloadable via the DAIC-WOZ project website at dcapswoz.ict.usc.edu. AVEC 2019 subset: 275 sessions (163 train, 56 dev, 56 test); includes audio, transcripts, visual and acoustic features; Accessed: 2025-01-30.

Mamadou Dia, Ghazaleh Khodabandelou, and Alice Othmani. 2024. Paying attention to uncertainty: A stochastic multimodal transformers for post-traumatic stress disorder detection using video. *Computer Methods and Programs in Biomedicine*, 257:108439.

Ming Fang, Siyu Peng, Yujia Liang, Chih-Cheng Hung, and Shuhua Liu. 2023a. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561.

Ming Fang, Siyu Peng, Yujia Liang, Chih-Cheng Hung, and Shuhua Liu. 2023b. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561.

Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, page 69–76, New York, NY, USA. Association for Computing Machinery.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yanrong Guo, Chenyang Zhu, Shijie Hao, and Richang Hong. 2022. A topic-attentive transformer-based model for multimodal depression detection. *arXiv preprint arXiv:2206.13256*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

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

Yonghong Li and Xiuzhuang Zhou. 2025. Fair uncertainty quantification for depression prediction. *arXiv preprint arXiv:2505.04931*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Aishik Mandal, Dana Atzil-Slonim, Thamar Solorio, and Iryna Gurevych. 2025. Enhancing depression detection via question-wise modality fusion. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 44–61, Albuquerque, New Mexico. Association for Computational Linguistics.

Kaining Mao, Wei Zhang, Deborah Baofeng Wang, Ang Li, Rongqi Jiao, Yanhui Zhu, Bin Wu, Tiansheng Zheng, Lei Qian, Wei Lyu, Minjie Ye, and Jie Chen. 2023. Prediction of depression severity based on the prosodic and semantic features with bidirectional lstm and time distributed cnn. *IEEE Transactions on Affective Computing*, 14(3):2251–2265.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Meng Niu, Kai Chen, Qingcai Chen, and Lufeng Yang. 2021. Hcag: A hierarchical context-aware graph attention model for depression detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4235–4239.

Mariia Nykoniuk, Oleh Basystiuk, Nataliya Shakhovska, and Nataliia Melnykova. 2025. Multimodal data fusion for depression detection approach. *Computation*, 13(1):9.

Syed Arbaaz Oureshi, Gaël Dias, Sriparna Saha, and Mohammed Hasanuzzaman. 2021. Gender-aware estimation of depression severity level in a multimodal setting. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Syed Arbaaz Qureshi, Mohammed Hasanuzzaman, Sriparna Saha, and Gaël Dias. 2019a. The verbal and non verbal signals of depression–combining acoustics, text and visuals for estimating depression level. *arXiv preprint arXiv:1904.07656*.

Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019b. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 81–88, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 55–63, New York, NY, USA. Association for Computing Machinery.

Morteza Rohanian, Julian Hough, and Matthew Purver. 2019. Detecting depression with word-level multimodal fusion. In *INTERSPEECH 2019*, pages 1443–1447.

Misha Sadeghi, Bernhard Egger, Reza Agahi, Robert Richer, Klara Capito, Lydia Helene Rupp, Lena Schindler-Gmelch, Matthias Berking, and Bjoern M. Eskofier. 2023. Exploring the capabilities of a language model-only approach for depression detection in text data. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5.

Misha Sadeghi, Robert Richer, Bernhard Egger, Lena Schindler-Gmelch, Lydia Helene Rupp, Farnaz Rahimi, Matthias Berking, and Bjoern M Eskofier. 2024. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, 3(1):66.

Evgeny A. Stepanov, Stéphane Lathuilière, Shammur Absar Chowdhury, Arindam Ghosh, Radu-Laurenţiu Vieriu, Nicu Sebe, and Giuseppe Riccardi. 2018. Depression severity estimation from multiple modalities. In *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

WHO. 2017. Depression and other common mental disorders: Global health estimates. Technical report, World Health Organization, Geneva. WHO/MSD/MER/2017.2.

WHO. 2022. World mental health report: Transforming mental health for all. Accessed: 2025-05-18.

James R. Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F. Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 11–18, New York, NY, USA. Association for Computing Machinery.

Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2017. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, page 53–59, New York, NY, USA. Association for Computing Machinery.

Jun Zhang and Yanrong Guo. 2024. Multilevel depression status detection based on fine-grained prompt learning. *Pattern Recogn. Lett.*, 178(C):167–173.

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

A PHQ-8 Depression Assessment

The Patient Health Questionnaire-8 (PHQ-8) (Kroenke et al., 2009) is a widely used self-report scale designed to measure the presence and severity of depressive symptoms. It is derived from the PHQ-9 but omits the ninth item concerning suicidal thoughts, making it more suitable for large-scale screening and automated processing.

Each of the eight items corresponds to a DSM-IV criterion for depression and asks respondents to rate how often they have experienced a specific symptom over the past two weeks. Responses are scored on a 4-point Likert scale:

- 0 – Not at all
- 1 – Several days
- 2 – More than half the days
- 3 – Nearly every day

The total PHQ-8 score ranges from 0 to 24 and is interpreted as follows:

- 0–4: None
- 5–9: Mild depression
- 10–14: Moderate depression
- 15–19: Moderately severe depression
- 20–24: Severe depression

The PHQ-8 has been validated in both clinical and general populations and is considered a reliable proxy for identifying depressive symptom severity in mental health research.

B Distress Analysis Interview Corpus (DAIC and E-DAIC)

The **Distress Analysis Interview Corpus** (DAIC-WOZ) (Gratch et al., 2014) and its extended version, **E-DAIC** (DAIC-WOZ Project, 2019), are widely used datasets for research in automated depression detection. Both datasets contain semi-structured clinical interviews conducted by a virtual interviewer named Ellie, operated via a "Wizard-of-Oz" setup, to elicit verbal and non-verbal indicators of psychological distress.

B.1 E-DAIC vs. DAIC-WOZ

The E-DAIC corpus is a re-transcribed and quality-controlled extension of DAIC-WOZ. It corrects known transcription errors and inconsistencies, and provides standardized splits for training, development, and testing. While DAIC-WOZ has been extensively used in prior work, E-DAIC offers improved data quality and is recommended for text-based modeling tasks.

B.2 Dataset Composition

E-DAIC consists of 275 participant interviews, partitioned as follows:

- **Training set**: 163 participants
- **Development set**: 56 participants

- **Test set**: 56 participants

Each session includes:

- **Audio recordings**: Interview audio in WAV format.
- **Transcripts**: Time-stamped dialogue with speaker labels.
- **Visual features**: Extracted using OpenFace, including facial landmarks, action units, and head pose.
- **Acoustic features**: Extracted via COVAREP and FORMANT analysis.
- **PHQ-8 scores**: Self-reported ratings of depression severity.

B.3 Data Organization

The dataset is organized into session-specific folders identified by participant IDs (e.g., `300_P`), each containing:

- `TRANSCRIPT.csv`: Annotated dialogue transcript.
- `AUDIO.wav`: Raw audio file.
- `COVAREP.csv`, `FORMANT.csv`: Acoustic features.
- `CLNF_features.txt`, `CLNF_AUs.csv`, `CLNF_pose.txt`, `CLNF_gaze.txt`: Visual features extracted using OpenFace.

Additional metadata includes:

- `train_split.csv`, `dev_split.csv`, `test_split.csv`: Partition definitions.
- `PHQ8_scores.csv`: Item-level and total PHQ-8 responses.

B.4 PHQ-8 Score Distribution

PHQ-8 scores in both DAIC and E-DAIC range from 0 to 24, capturing varying levels of depressive symptom severity. The distribution is right-skewed, with a concentration of low-to-moderate severity cases, which presents challenges for model calibration and minority class performance.

B.5 Usage Considerations

Researchers working with DAIC or E-DAIC should consider the following:

- **Data Quality**: E-DAIC addresses known issues in DAIC-WOZ, including transcript errors and missing data, and is recommended for textual modeling.
- **Ethical Use**: Given the sensitive nature of the interviews, ethical guidelines and approvals must be followed.
- **Licensing**: Access requires agreement to the dataset's End User License Agreement (EULA).

Our use of both datasets complies with their intended research purpose. The corpora were released to support research on automated detection of psychological distress and related mental health conditions. In this work, we focus exclusively on the prediction of PHQ-8 depression severity from textual transcripts, a primary task for which the dataset was designed. The datasets are anonymized at source, with personally identifiable information removed prior to distribution. We further restrict our usage to non-commercial, academic settings, operate solely on de-identified utterance sequences, and report only aggregate results. No individual-level data or metadata are released. All use complies with the dataset's End User License Agreement (EULA) and contributes to its intended goal of advancing computational methods for mental health assessment.

For detailed information on data preprocessing and feature extraction methodologies, refer to the official documentation provided with the dataset.

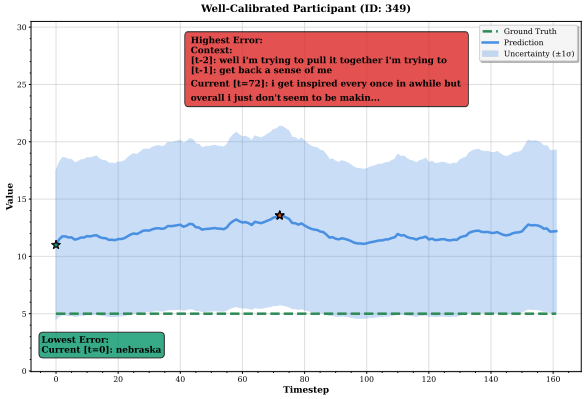
C Ablation Study Experimental Setup

For each ablation, we use the same data splits, batch size, optimizer, learning rate schedule, and early stopping criteria as the main experiments. The following configurations are evaluated:

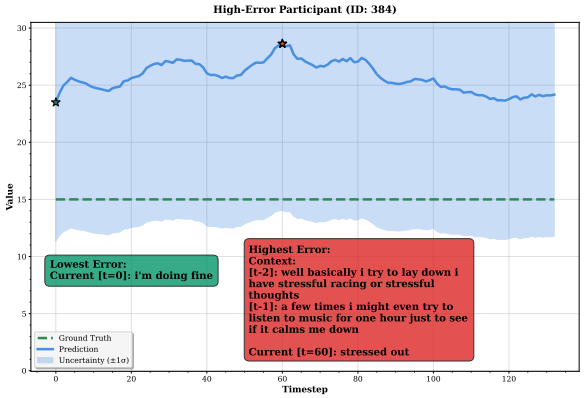
- **Full Model**: All components enabled (attention, residual, variance).
- **No Attention**: Attention layer removed.
- **No Residual**: Residual connection removed.
- **No Variance**: Variance prediction head disabled; model trained with MSE loss.

Each model is trained for the same number of epochs with fixed random seeds for reproducibility. After training, we evaluate on the held-out test set and report MAE, RMSE, and NLL (where available). All code, configurations, and results are available for reproducibility.

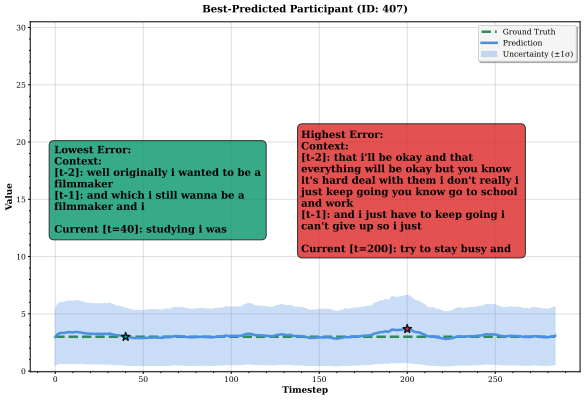
D Case Study



(a) well-calibrated

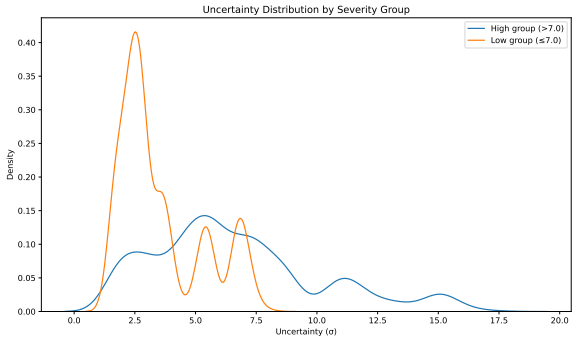


(b) high-error, high-uncertainty

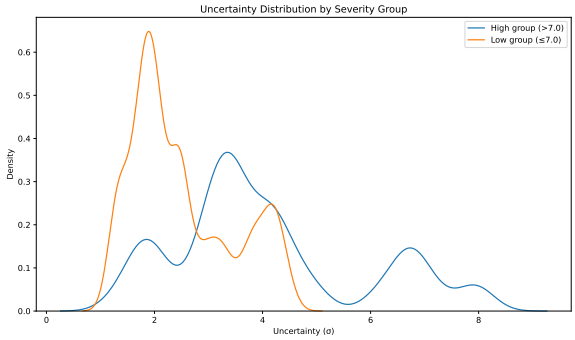


(c) best predicted

Figure 4: Case studies illustrating prediction, ground truth, and model uncertainty for three representative participants. Each subplot shows the temporal trajectory of predictions (blue), ground truth (dashed green), and uncertainty intervals (light blue area) across timesteps.



(a) $\alpha = 1, \beta = 1, \gamma = 1$



(b) $\alpha = 1, \beta = 1, \gamma = 0.5$

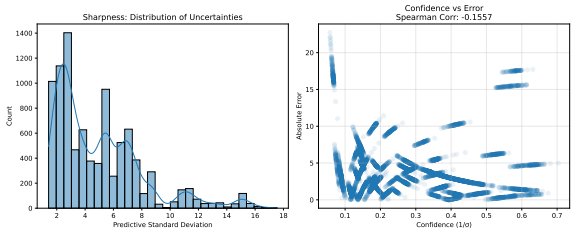
Figure 5: Uncertainty distributions by severity group

To provide a more detailed understanding of the model's behavior, we conducted case studies on selected participants, focusing on key scenarios such as well-calibrated predictions, high uncertainty, high error, and best-predicted cases. For each participant, we analyzed their prediction trajectories, uncertainty estimates, and ground truth values over time. Figure 4 illustrates three representative examples: (a) a well-calibrated participant where predicted uncertainties closely align with observed errors; (b) a high-error, high-uncertainty case, reflecting model uncertainty under ambiguous input; and (c) the best-predicted participant, demonstrating accurate predictions with narrow uncertainty bands. These examples highlight the model's ability to adaptively express confidence, offering interpretable outputs for both reliable and uncertain predictions.

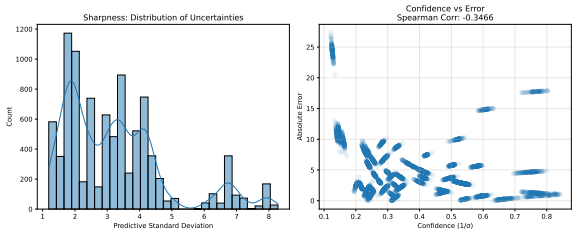
E Further experiments

E.1 Sharpness Calibration Tradeoff.

To further analyze the quality of our uncertainty estimates, we examine the sharpness–calibration tradeoff. Sharpness refers to the concentration or narrowness of the model's predictive distributions, with sharper (lower variance) predictions indicat-



(a) $\alpha = 1, \beta = 1, \gamma = 1$



(b) $\alpha = 1, \beta = 1, \gamma = 0.5$

Figure 6: Sharpness calibration tradeoff

ing higher confidence. However, sharpness must be balanced with calibration: a model that is too sharp may be overconfident, while a model that is too broad may be underconfident. Figure 6 visualizes the distribution of predictive standard deviations across the test set and assesses the relationship between predicted uncertainty and actual error. This analysis reveals whether the model's most confident predictions are indeed more accurate, and whether improvements in sharpness come at the expense of calibration.

We observe that the model with $\gamma = 0.5$ produces a sharper distribution of predictive standard deviations, reflecting lower predicted uncertainty overall. This configuration also yields a stronger negative correlation between predicted standard deviation and absolute error ($r = -0.3466$), compared to the default uniform configuration ($r = -0.1557$). This indicates that, under $\gamma = 0.5$, the model's uncertainty estimates more effectively distinguish between high- and low-error predictions. However, as discussed previously, this gain in sharpness and ranking quality comes at the cost of calibration: the model systematically underestimates its uncertainty, leading to undercoverage in the prediction interval analysis.

E.2 Error–Uncertainty Correlation

Figure 7 illustrates the relationship between predicted uncertainty and absolute prediction error for PTTSD trained with Gaussian NLL. We observe a strong linear correlation, with a Pearson coefficient of 0.88 and Spearman rank correlation of 0.64, both statistically significant ($p < 0.001$). A fitted regression line yields an R^2 of 0.77 with a narrow 95% confidence interval, confirming that higher uncertainty estimates are predictive of higher errors. This supports the model's ability to assign meaningful and interpretable uncertainty in practice.

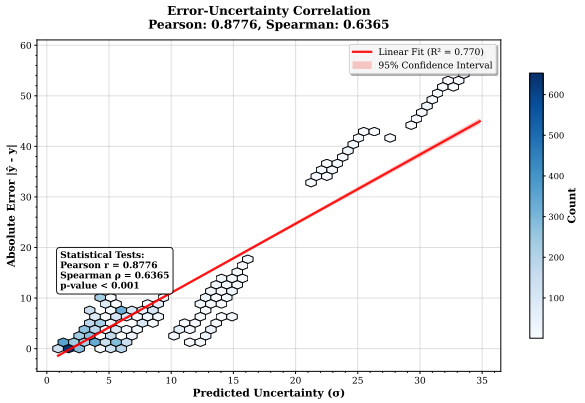


Figure 7: Error uncertainty correlation for PTTSD with Gaussian NLL

E.3 Temporal Uncertainty Dynamics

Figure 8 shows how mean predicted uncertainty and mean absolute error evolve across timesteps for PTTSD trained with Gaussian NLL. Early in the sequence, where many samples are available, both uncertainty and error are relatively high but decrease steadily as the model accumulates contextual information. After around timestep 250, uncertainty stabilizes, while the error begins to increase. This divergence is likely due to the sharp drop in sample count at later timesteps (e.g., only 8 samples after timestep 200 and just one after timestep 350), which introduces statistical noise and limits the model's ability to generalize.

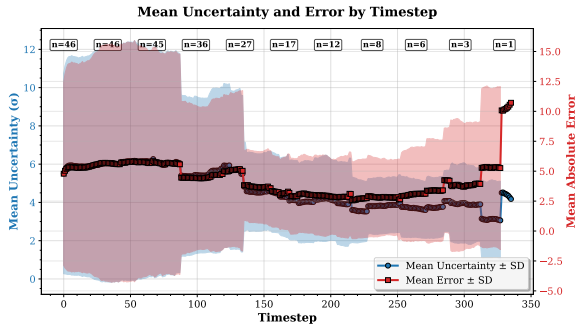


Figure 8: Temporal uncertainty tendencies for PTTSD with Gaussian NLL