Probabilistic Textual Time Series Depression Detection

Anonymous ACL submission

Abstract

We propose PTTSD, a *Probabilistic Textual Time Series Depression Detection* framework for predicting PHQ-8 depression severity scores from utterance-level clinical interviews. PTTSD models both predictive means and calibrated uncertainty over time using Gaussian and Student's-t distributions, trained via negative log-likelihood losses. Our architecture combines bidirectional LSTMs with self-attention and residual connections to model textual sequences, and employs uncertainty-aware output heads for calibrated probabilistic predictions. On the E-DAIC dataset, PTTSD achieves state-of-the-art performance among text-only systems (MAE = 3.85, RMSE = 4.52), outperforming recent baselines. Extensive ablation and sensitivity studies underscore the value of self-attention, probabilistic modeling, and calibrated uncertainty, establishing PTTSD as a robust and interpretable framework for uncertainty-aware depression forecasting in clinical NLP.

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 can complement traditional screening, especially in digital therapy and remote care contexts. Text-based systems that model clinical interviews have shown promise for predicting standardized scores such as the PHQ-8.

Recent work on textual depression detection has focused on deterministic models—LSTMs, Transformers, or large language models (LLMs) (Mandal et al., 2025; Fang et al., 2023; Nykoniuk et al., 2025; Sadeghi et al., 2024)—that output scalar severity estimates from utterance sequences. While effective at sequence modeling, these approaches provide point predictions without quantifying uncertainty, limiting their interpretability and reliability in sensitive domains like mental health care.

We introduce PTTSD, a *Probabilistic Textual Time Series Depression Detection* framework that addresses this limitation. PTTSD models utterance-level textual sequences using a probabilistic LSTM with self-attention and residual connections, and produces calibrated uncertainty estimates via Gaussian or Student's-t output distributions. It is trained with negative log-likelihood losses, enabling distributional predictions rather than point estimates.

We evaluate PTTSD on the E-DAIC dataset and demonstrate strong results across PHQ-8 prediction metrics (MAE, RMSE), outperforming recent text-based systems requiring no handcrafted features or prompt engineering. In-depth ablation studies and calibration analysis reveal the model's sensitivity to loss design and architectural components. PTTSD combines strong empirical performance with uncertainty awareness, offering a robust building block for mental health NLP applications.

While prior work has made substantial progress in text-based depression detection, several key limitations remain. First, most existing approaches rely on deterministic models that provide point predictions without expressing confidence or uncertainty, making them ill-suited for risk-sensitive clinical settings. Second, prompt-based systems such as those proposed by Sadeghi et al. (2024) require extensive experimentation with prompt variants and post hoc selection, increasing complexity and reducing reproducibility. Third, several models process only a subset of the available interview utterances (e.g., question-response pairs), potentially discarding valuable temporal information distributed across the full conversation.

Our contributions are as follows:

- We propose PTTSD, a fully probabilistic sequence model that jointly predicts PHQ-8 scores and calibrated uncertainty from

utterance-level text sequences using Gaussian and Student's-t output distributions.

- We train and evaluate PTTSD end-to-end on all available utterances without handcrafted prompt design or selection, providing a simple and reproducible modeling pipeline.
- We demonstrate state-of-the-art results on the E-DAIC benchmark among fully automatic, text-only systems, and conduct extensive ablations, sensitivity analysis, and calibration evaluations to understand uncertainty quality and model robustness.

The remainder of this paper is structured as follows. In Section 2, we review prior work on depression detection from text and highlight the gap in probabilistic modeling. Section 3 introduces the PTTSD architecture, including model components, probabilistic loss functions, and training procedures. Section 4 presents our experimental setup and main results, including comparisons to baselines, ablation studies, and uncertainty calibration analysis. Finally, Section 5 concludes with a discussion of limitations and directions for future work.

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., 2023; 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. (2019) 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. (2022) 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 Data and Preprocessing

We utilize the Extended Distress Analysis Interview Corpus (E-DAIC) (Gratch et al., 2014), which contains 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 (Patient 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 E-DAIC in Appendix A and Appendix B, respectively.

To improve interview transcription fidelity, we

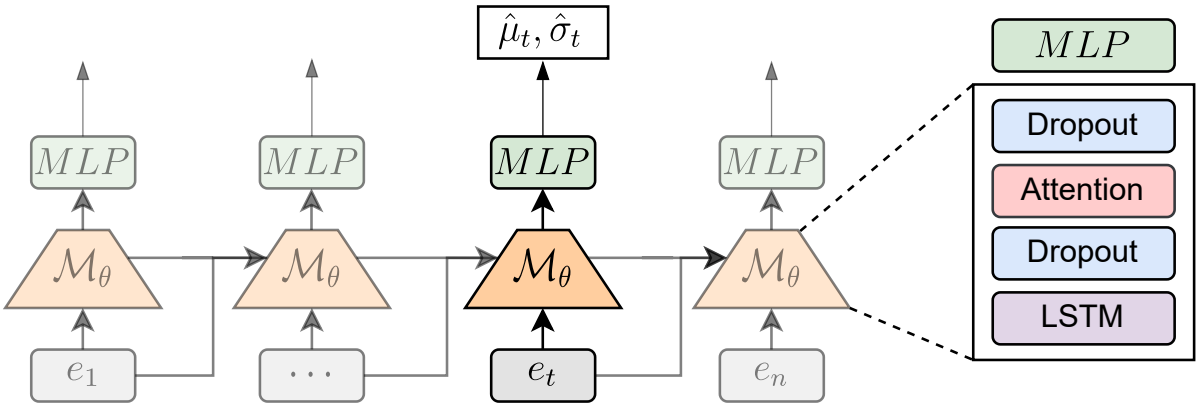


Figure 1: Probabilistic Textual Time Series Depression Detection

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 in (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.

3.2 Generating Utterance Embeddings

We represent each utterance using the `all-MiniLM-L6-v2`[1] Sentence Transformer (Reimers and Gurevych, 2019), which we found to outperform other tested embedding models (e.g., standard BERT) in preliminary experiments. Each utterance is independently encoded into a fixed-dimensional vector ($e_t \in \mathbb{R}^D$ using a pretrained language model. The resulting embedding sequence is $(e_1, e_2, \ldots, e_T)$, where T is the utterance sequence length. These embeddings form the input to the model $\mathcal{M}$. Utterance embeddings are stacked into a tensor $\mathbf{X} \in \mathbb{R}^{B\times T\times D}$, where B denotes the batch size, T the number of utterances per sequence, and D the dimensionality of each embedding. Attention masks are propagated throughout the pipeline to mask out padded positions during modeling, loss computation, and evaluation.

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

3.3 Probabilistic LSTM Sequence-to-Sequence

Inspired by the architecture of Mandal et al. (2025), we first encode $\mathbf{X}$ using a multi-layer bidirectional LSTM. The resulting hidden sequence $\mathbf{H} \in \mathbb{R}^{B\times T\times H}$ is passed through a multi-head self-attention layer (Vaswani et al., 2017) to capture long-range dependencies. A residual connection is applied between the LSTM and attention outputs. Two feedforward networks then predict the mean $\hat{\mu}_t$ and standard deviation $\hat{\sigma}_t$ at each time step:

$$\hat{\mu}_t = f_{\text{mean}}(e_t), \qquad \hat{\sigma}_t = \text{softplus}(f_{\text{std}}(e_t)) + \epsilon$$

Dropout is applied after the LSTM and within the MLPs. All predictions and ground truth values are masked to select only valid, non-padded positions. An overview of this architecture is illustrated in Figure 1.

3.4 Sequence Modeling and Predictive Distributions

We model the PHQ-8 score 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 in parallel across all time steps (i.e., non-autoregressively), and does not receive ground truth labels $y_{<t}$ or past predictions.

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:

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

3.5 Loss Functions

The total sequence loss is the negative log-likelihood of all valid time steps:

$$\mathcal{L}_{\text{seq}} = -\sum_{t=1}^{T} \log p(y_t \mid e_{\leq t}; \theta)$$

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 valid sequence length for participant i. When using Gaussian outputs, the loss becomes:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{2N} \sum_{i=1}^{N} \left[\log(2\pi\hat{\sigma}_i^2) + \left(\frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i}\right)^2\right]$$

We optionally use MSE or MAE as auxiliary losses for ablation.

3.6 Training Procedure

PTTSD is trained for 50 epochs using the Adam optimizer with a constant learning rate (2e-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., 2017). 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 (163, 56, and 56 samples, respectively) provided in the E-DAIC dataset. As described in Section 3.1, 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.[2] To account for variability due to random initialization, we report average performance over three runs with different seeds.

4.2 Main Results

Table 1 reports the performance of our proposed model PTTSD alongside a range of text-only baselines for PHQ-8 prediction on the E-DAIC dataset. PTTSD achieves the lowest test MAE (3.85) and RMSE (4.52), setting a new state of the art among fully automated, text-based systems.

Early approaches such as the LSTM-based multi-level attention network from Ray et al. (2019) and the CNN-LSTM variants by Rodrigues Makiuchi et al. (2019) demonstrate competitive but overall lower performance, with test RMSEs of 4.73 and 6.88, respectively. While Rodrigues Makiuchi et al. (2019) reports a stronger dev RMSE (4.22) using 8 CNN blocks, no corresponding test results are provided for that setting, limiting direct comparability.

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

Method	MAE (Dev)	RMSE (Dev)	MAE (Test)	RMSE (Test)
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 (ours)	3.47 (±0.017)	4.57 (±0.041)	**3.85** (±0.041)	**4.52** (±0.38)

Table 1: Evaluation of PHQ-8 regression performance across text-only models on the E-DAIC dataset. Results for related work are taken from Sadeghi et al. (2024). Bold results indicate best performance.

Recent works by Sadeghi et al. (2023, 2024) leverage prompt-based large language models and Whisper-based transcriptions. Among these, the Pr3+Whisper variant performs best (test MAE 4.22, RMSE 5.07), while the top dev results are achieved by Pr3+Whisper+AudioQual (MAE 2.85, RMSE 4.02). However, this latter model involves audio-based quality filtering and is not strictly text-only, making PTTSD the best-performing model under the text-only constraint.

The strength of PTTSD lies not only in its empirical performance but also in its simplicity and generalizability. Unlike prior work such as Sadeghi et al. (2024), which evaluates multiple prompt variants and selects the best-performing configuration post hoc, PTTSD trains and evaluates a single, unified model architecture end-to-end. This eliminates the need for prompt engineering.

4.3 Ablation Studies

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: Comparison of loss functions on development and test sets.

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 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.

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.

Table 3 and Figure 2 illustrate the effects of disabling different components. Removing self-attention yields the largest degradation in performance, 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 no-

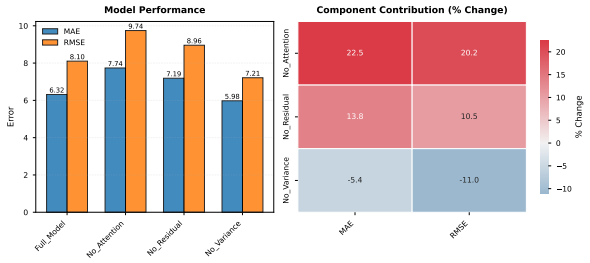


Figure 2: Ablation results

ticeable 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 results in better MAE and RMSE (–5.4% and –11.0%, respectively), likely due to the simpler deterministic regression objective. However, this simplification eliminates the model's ability to quantify uncertainty—an essential capability in risk-sensitive applications like mental health prediction.

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.

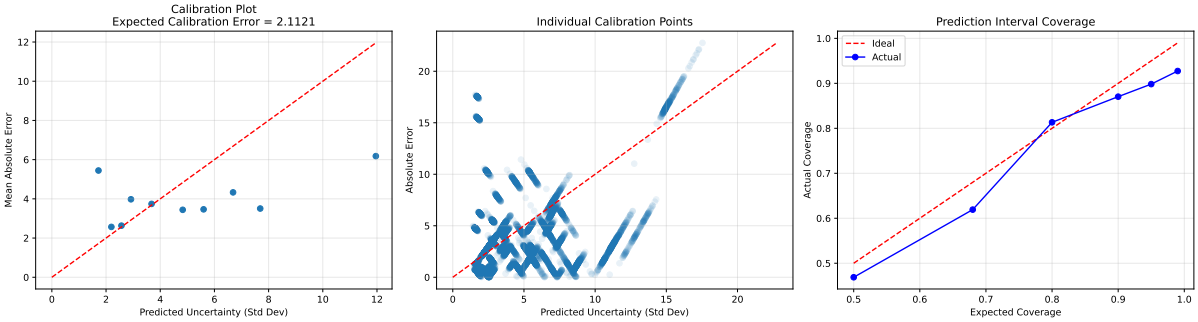
Table 4 presents the effect of varying the NLL weighting parameters β (log-variance term) and γ (normalized squared error term), with α held constant as it weights the constant term in the NLL and hence does not influence the model's gradients or learning dynamics. The standard setting ($\beta = \gamma = 1$) 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 Analysis

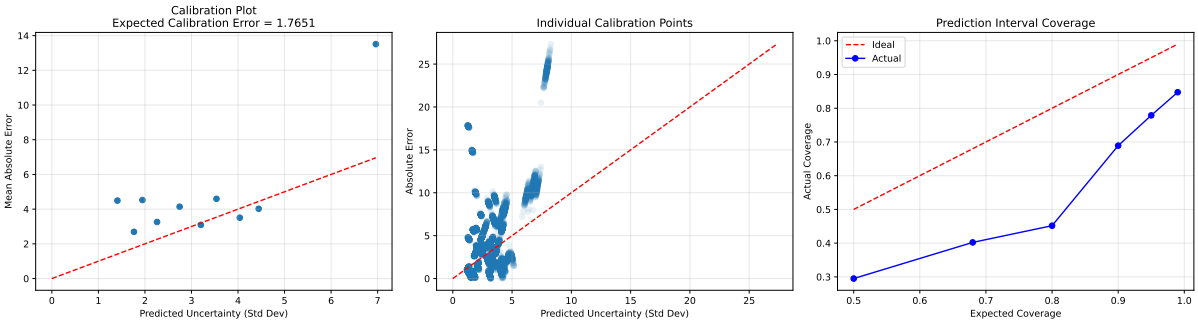
Calibration Analysis. To evaluate the quality of our model's uncertainty estimates, we conduct a three-part calibration analysis shown in Figure 3. First, the binned calibration plot (left) groups predictions by predicted uncertainty and compares the mean predicted standard deviation (x-axis) with the mean absolute error (y-axis) in each bin. Perfect calibration lies on the red diagonal, with deviations quantified by the Expected Calibration Error (ECE). Next, the individual calibration plot (middle) displays each test prediction as a scatter point, with predicted uncertainty on the x-axis and the observed absolute error on the y-axis. This view provides fine-grained insight into the relationship between uncertainty and error across instances. Finally, the coverage plot (right) evaluates the proportion of ground truth values falling within the model's prediction intervals at various confidence levels. Ideal calibration lies on the red diagonal; deviations above or below reflect under- or over-confident interval estimates, respectively.

We visualize calibration results for Gaussian NLL under two hyperparameter settings. The first uses the standard configuration $\alpha = 1, \beta = 1, \gamma = 1$, which achieved the best overall performance (Table 1) and is shown in Figure 3a. The second setting, shown in Figure 3b, prioritizes calibration by reducing γ to 0.5. As discussed earlier, this leads to worse NLL on the development and test sets, but improves calibration—evident in the left and middle plots of Figure 3, as well as in the lower Expected Calibration Error (2.1121 vs. 1.7651). However, the model becomes overconfident, as indicated by the coverage plot falling below the ideal diagonal, meaning it underestimates its predictive uncertainty.

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



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

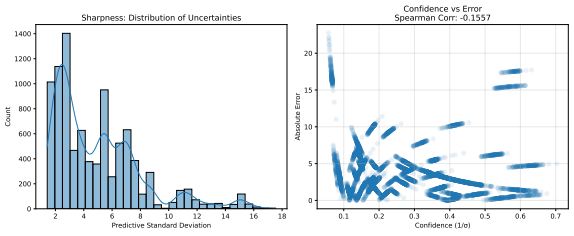


(b) $\alpha = 1, \beta = 1, \gamma = 0.5$

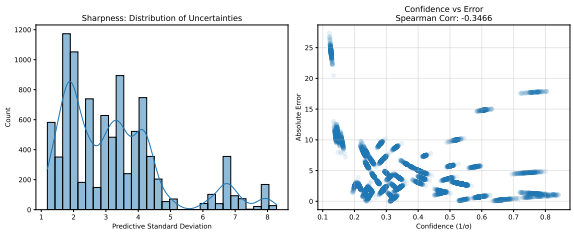
Figure 3: Calibration analysis of the predicted uncertainties for Gaussian NLL on the test set

the model's predictive distributions, with sharper (lower variance) predictions indicating 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 4 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.



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



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

Figure 4: Sharpness Calibration Tradeoff

5 Conclusion

We introduced PTTSD, a novel probabilistic neural framework for predicting PHQ-8 depression severity from utterance-level textual sequences. Unlike prior work that outputs deterministic point estimates, PTTSD models calibrated predictive uncertainty using Gaussian and Student's-t distributions. Our architecture combines bidirectional LSTMs, self-attention, and residual connections, and is trained via negative log-likelihood losses. PTTSD is fully data-driven and requires no manual feature engineering or prompt-based supervision, enhancing its applicability in real-world clinical settings where manual intervention is infeasible. Empirical evaluation on the E-DAIC dataset shows that PTTSD achieves state-of-the-art performance among fully automatic, text-only systems, outperforming recent baselines. Ablation studies confirm the value of attention and probabilistic heads, while sensitivity analysis highlights the importance of balanced loss weighting. Calibration analysis further supports the reliability of PTTSD's uncertainty estimates. The results demonstrate that uncertainty-aware textual time series modeling is both feasible and beneficial for clinical NLP. Future work will extend PTTSD to multimodal inputs and investigate its deployment in real-world digital mental health tools.

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

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.

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. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561.

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).

Jonathan Gratch and 1 others. 2020. Extended distress analysis interview corpus (e-daic). https:

//dcapswoz.ict.usc.edu/. Accessed: 2025-01-30.

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*.

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*.

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, and 1 others. 2022. 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.

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

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. 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.

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.

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.

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 Extended Distress Analysis Interview Corpus (E-DAIC)

The Extended Distress Analysis Interview Corpus (E-DAIC) (Gratch et al., 2020) is an enriched version of the DAIC-WOZ dataset (Gratch et al., 2014), designed to facilitate research in automated depression detection. It comprises semi-structured interviews conducted by a virtual interviewer named Ellie, controlled by a human operator in a "Wizard-of-Oz" setup. These interviews aim to elicit verbal and non-verbal indicators of psychological distress.

B.1 Dataset Composition

E-DAIC includes data from 275 participants, with the following partitioning:

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

Each participant's session contains:

- **Audio recordings**: Captured in WAV format.
- **Transcripts**: Annotated with time stamps and speaker labels.
- **Visual features**: Extracted using tools like OpenFace, including facial landmarks and action units.
- **Acoustic features**: Derived using COVAREP, encompassing prosodic and voice quality metrics.
- **PHQ-8 scores**: Self-reported assessments of depression severity.

B.2 Data Organization

The dataset is organized into session-specific folders named with participant IDs (e.g., `300_P`). Each folder contains:

- `TRANSCRIPT.csv`: Dialogue transcripts with time-aligned annotations.
- `AUDIO.wav`: Raw audio recordings of the interview.
- `COVAREP.csv`: Acoustic feature sets.
- `FORMANT.csv`: Formant frequency features.
- `CLNF_features.txt`: 2D facial landmark positions.
- `CLNF_AUs.csv`: Facial Action Units data.
- `CLNF_gaze.txt`: Gaze tracking information.
- `CLNF_pose.txt`: Head pose estimations.

Additionally, the dataset includes metadata files:

- `train_split.csv`, `dev_split.csv`, `test_split.csv`: Define the dataset partitions.
- `PHQ8_scores.csv`: Contains individual item responses and total scores.

B.3 PHQ-8 Score Distribution

The PHQ-8 scores in E-DAIC range from 0 to 24, reflecting varying levels of depression severity. The distribution is skewed towards lower scores, indicating a higher number of participants with minimal depressive symptoms. This imbalance poses challenges for training models to accurately predict higher severity levels.

B.4 Usage Considerations

Researchers utilizing E-DAIC should be aware of certain factors:

- **Data Quality**: Some sessions may have missing or incomplete data due to technical issues during recording.
- **Ethical Use**: As the dataset involves sensitive mental health information, appropriate ethical considerations and approvals are necessary for its use.
- **Licensing**: Access to E-DAIC requires agreement to a specific End User License Agreement (EULA) set by the data providers.

Our use of the E-DAIC dataset is fully consistent with its intended purpose. The corpus was 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 dataset is 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.