Detecting PTSD in Clinical Interviews: A Comparative Analysis of NLP Methods and Large Language Models

Feng Chen[1,2], Dror Ben-Zeev[2], Gillian Sparks[2], Arya Kadakia[2], Trevor Cohen[1,2]
[1] Department of Biomedical Informatics and Health Education, University of Washington
[2] BRiTE Center, Department of Psychiatry and Behavioral Sciences, University of Washington

BACKGROUND

- PTSD arises after trauma and is often under-diagnosed, leading to chronic suffering and impaired functioning.
- Around 6% of adults in the U.S. will experience PTSD in their lifetime.[1]
- Routine screening in primary care could improve detection and enable timely therapeutic interventions.

DATASET

- DAIC-WOZ clinical interview transcripts (189 participants).[2,3]
- 56 participants (29.6%) were diagnosed with PTSD based on the PTSD Checklist–Civilian Version (PCL-C); 133 non-PTSD cases.
- 42 participants (22.2%) also screened positive for depression using the Patient Health Questionnaire-8 (PHQ-8).
- AUPRC used as primary metric due to class imbalance.

METHODS

Fine-Tuning
BERT-base,
RoBERTa-base
Mental-BERT,
Mental-RoBERTa

Embeddings + LR
SentenceBERT + LR
LLaMA embeddings + LR

LLM Prompting
Zero-shot (DSM-5)
Few-shot (examples)
Chain-of-thought

RESULTS

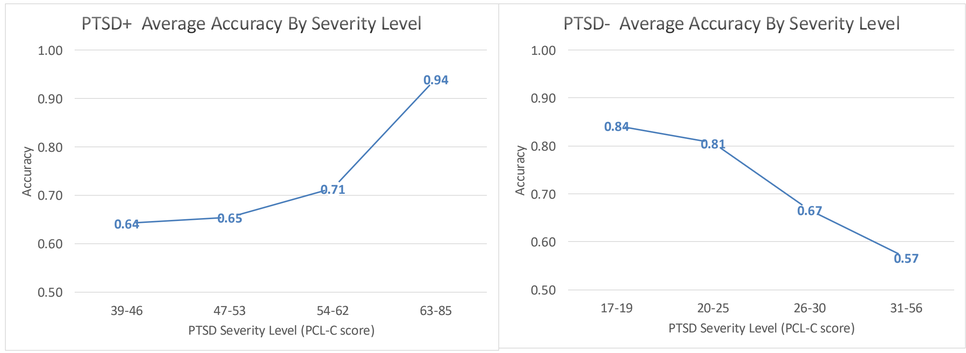
Performance for BERT-based models

	BERT-base	Mental-BERT	RoBERTa-base	Mental-RoBERTa
AUPRC	0.522±0.0 92	0.647±0.101	0.599±0.145	**0.675±0.084**
AUC	0.723±0.108	**0.794±0.059**	0.723±0.107	0.786±0.046
Balanced Acc	0.702±0.101	**0.739±0.091**	0.662±0.088	0.695±0.092

Cross-approach Performance

	Best BERT-based model	Embedding-based model		Prompt-based model		
	Mental RoBERTa	SBERT + LR	LLaMA + LR	LLaMA ZS	LLaMA FS	LLaMA CoT
AUPRC	0.675±0.084	**0.758±0.128**	0.693±0.094	0.701	0.737	0.579
AUC	0.786±0.046	0.856±0.069	0.835±0.046	0.841	**0.875**	0.705
Balanced Acc	0.695±0.092	0.801±0.097	0.747±0.029	0.751	**0.804**	0.681

Severity & Comorbidity



AUPRC	patient with depression	Patient without depression
Mental RoBERTa	0.953±0.048	0.521±0.171
SBERT + LR	0.925±0.073	**0.574±0.137**
LLaMA + LR	0.879±0.070	0.562±0.100
LLaMA zs	0.915	0.492
LLaMA fs	**0.961**	0.467
LLaMA cot	0.859	0.381

DISCUSSION

- SentenceBERT + LR outperforms fine-tuning approaches; domain-specific pre-training (e.g., Mental-RoBERTa) yields moderate gains.
- Zero-/few-shot prompts provide competitive performance while chain-of-thought underperforms.
- Accuracy improves with PTSD severity but declines for non-PTSD; comorbid depression shows high AUPRC (~0.95 vs ~0.50).
- Embedding-based representations of clinical text are robust, offering scalable options for screening.
- Moderate-severity PTSD cases remain difficult to detect, highlighting a critical clinical challenge.
- Comorbidity with depression provides cross-symptom cues that can aid detection in some cases.

CONCLUSIONS

- SentenceBERT + LR attains the highest AUPRC (0.758) with computational efficiency; Mental-RoBERTa excels among fine-tuned models.
- Few-shot prompts deliver strong discrimination (AUROC 0.875) without labelled data; zero-shot remains competitive.
- Severe PTSD and comorbid depression can be detected reliably, but moderate-severity and non-depressed cases remain challenging.

ACKNOWLEDGEMENTS

The authors of this study are supported by a grant from the National Institute of Mental Health (# U01MH135901). We would like to thank John Gratch, Jill Boberg, and their colleagues at the USC Institute for Creative Technologies for making this dataset available to the scientific community.

[1]: Affairs USDoV. How Common Is PTSD in Adults? National Center for PTSD2025 [Available from: https://www.ptsd.va.gov/understand/common/common_adults.asp.
[2]: Gratch J, Artstein R, Lucas GM, et al., editors. The distress analysis interview corpus of human and computer interviews. LREC; 2014: Reykjavik.
[3]: DeVault D, Artstein R, Benn G, et al., editors. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems; 2014.