# AUTOMATING PTSD DIAGNOSTICS IN CLINICAL INTERVIEWS: LEVERAGING LARGE LANGUAGE MODELS FOR TRAUMA ASSESSMENTS

Sichang Tu, Abigail Powers, Natalie Merrill, Negar Fani, Sierra Carter, Stephen Doogan, Jinho D. Choi

## BACKGROUND

**Service Gap** The scarce clinical workforce presents significant challenges in mental healthcare, limiting access to formal diagnostics and services.

**Research Gap** No work has employed diagnostic interviews between real clinicians and patients that are systematically conducted.

**Goal** Automates PTSD diagnostic assessments based on structured clinician-administered interviews.

## PTSD INTERVIEW DATA

This study utilizes data from 411 clinician-administrated diagnostic interviews conducted with 336 participants from a larger study on risk resiliency to the PTSD development in a population seeking medical care.

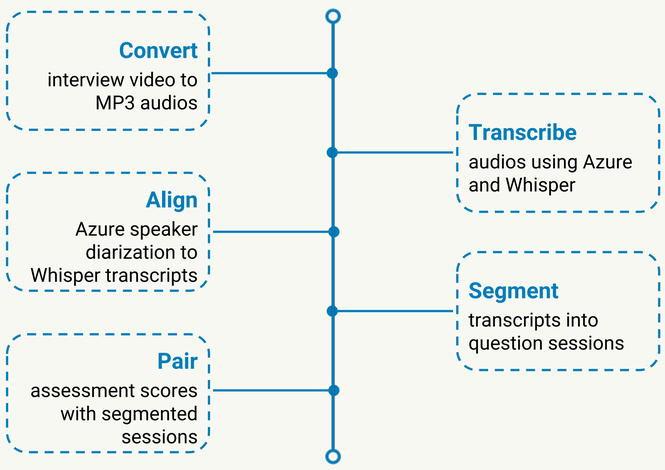
We focus on 4 out of 10 sections which are applied to the majority of participants. These include the internally designed Life Base Interview (`LBI`) and Treatment History & Health (`THH`), for accessing psychiatric history, treatment, and suicidality, alongside the Criterion A (`CRA`) and the Clinician-Administered PTSD Scale for DSM-5 (`CAP`), which adhere to standard PTSD criteria in Diagnostic and Statistical Manual of Mental Disorders.

| Sec | Ques | Vars | Example Question | Exam. Var |
|---|---|---|---|---|
| `LBI` | 31 | 15 | What has been your primary source of income over the past month? | lbi_a1 |
| `THH` | 39 | 20 | In the past, have you been treated for any emotional/mental health problems with therapy or hospitalization? | thh_tx_yesno |
| `CRA` | 17 | 20 | What would you say is the one that has been most impactful where you are still noticing it affecting you? | critaprobenotes |
| `CAP` | 241 | 92 | In the past month, have you had any unwanted memories of the [Event] while you were awake, so not counting dreams? | dsm5capscritb01<br>trauma1_distress |

*Statistics and examples for each of the four sections employed in this study.*

## DATA PROCESSING

Each interview video is processed following the pipeline:




| Type | Variables | | | | | Count |
|---|---|---|---|---|---|---|
| | LBI | THH | CRA | CAP | Total | |
| `Scale` | 7 | 1 | 0 | 40 | 48 | 9,722 |
| `Category` | 4 | 9 | 15 | 3 | 31 | 4,258 |
| `Measure` | 2 | 0 | 1 | 24 | 27 | 3,482 |
| `Notes` | 1 | 10 | 3 | 0 | 14 | 1,146 |
| `Rule` | 1 | 0 | 1 | 25 | 27 | 6,326 |

**Variable statistics**
Our data comprises 5 variable types, corresponding to different value types of interview answers.

**Dataset statistics**
The first and largest PTSD clinician-administered interview dataset.

| | Audios | Hours | Turns | Token |
|---|---|---|---|---|
| Original | 456 | 779 | 260,688 | 6,656,092 |
| Transcribe | 435 | 721 | 200,545 | 6,071,867 |
| Evaluation | 322 | 512 | 142,824 | 4,335,977 |

## EXPERIMENTS

- 2 state-of-the-art Large Language Models: GPT-2 and Llama-2.
- Develop prompt template for each variable type with replaceable patterns.
- Experiment with zero-shot and few-shot settings for both models.

| VT | Template |
|---|---|
| **S&C** | `[INTRO].` Based on the patient's interview history, please determine *{keywords}* that the patient *{symptom}*. `[RETURN]. [REASON].` The "answer" should be in the range {range}.{attributes} |
| **M** | `[INTRO].` Based on the patient's interview history, please calculate *{keywords}* that the patient have *{symptom}*. `[RETURN]. [REASON].` The "answer" should be *{type}*. |
| **N** | `[INTRO].` Based on the formatted data from patient's interview, please determine whether or not the formatted data includes this specified information *{single_slot}*. `[RETURN].` The "reason" gives a brief explanation on whether the formatted data includes or omits the information. The "answer" should be either "yes" or "no", indicating the presence or absence of the information in formatted data. |

*Instruction templates for **S**cale, **C**ategory, **M**easure, and **N**otes variables. VT: Variable type, `[INTRO]`: Imagine you are a professional clinician, `[RETURN]`: Return the answer as a JSON object with "reason" and "answer" as the keys, `[REASON]`: The "reason" should provide a brief justification or explanation for the answer.*

## RESULTS

We adopt 4 evaluation metrics for different variable types: Accuracy, Recall, Root Mean Square Error (RMSE), and Bias evaluation.

| Type | Count | Acc | | RMSE | | Bias | | Recall | |
|---|---|---|---|---|---|---|---|---|---|
| | | GPT4 | Llama2 | GPT4 | Llama2 | GPT4 | Llama2 | GPT4 | Llama2 |
| `Scale` | 9,722 | 58.9 | 46.7 | 1.10 | 1.63 | -0.04 | 0.51 | - | - |
| $\texttt{Scale}_g$ | 9,722 | 67.3 | 59.0 | 0.85 | 1.01 | -0.04 | 0.51 | - | - |
| `Category` | 4,258 | 77.2 | 63.6 | - | - | - | - | - | - |
| `Measure` | 3,482 | 64.4 | 56.5 | - | - | -0.34 | -0.004 | - | - |
| `Notes` | 1,146 | - | - | - | - | - | - | 48.1 | 52.7 |
| `Rule` | 6,326 | 68.4 | 59.8 | 0.80 | 0.92 | -0.15 | 0.44 | - | - |

*Model performance on all variable types using four evaluation metrics.*

- GPT achieves 10.5% higher performance than Llama on average.
- Llama exhibits better performance with a recall of 52.7% than GPT for Notes.
- GPT and Llama achieve RMSE < 1 for Rule (predictions are less than one scale off from gold).
- GPT displays a marginal bias toward negative for Scale, while Llama shows a strong positive bias.

## ERROR ANALYSIS

**Misaligned Reasoning** Models deviate from instructions of the rating scheme, presenting seemingly logical reasoning, although it ultimately leads to incorrect conclusions.

**False Negatives** 1. Inaccurate assessments by clinicians; 2. Ambiguity in Scale variables where answers may fall between two scales, 3. The model's inability to recognize paraphrased information in Notes variables.

**External Information** Errors are caused by the absence of external information, such as the prior knowledge about the patient (e.g., medical histories, demographics) or the content of previous interview questions.

**Transcription Error** Errors from automatic speech recognizers often cause LLMs to incorrectly interpret the answers, especially with short responses (e.g., yes, no, single digits like 6), medical terminologies, or non-verbal cues such as nodding.

**Session Mismatching** The segmented session may not contain all the necessary information due to mismatched question span, especially when the clinician extensively paraphrases it.

**Commonsense Reasoning** The model fails to infer ordinary human experiences and routines.

**Q:** the intensity of physical reactions in the past month.

**History:**
*Have you had any physical reactions when something reminded you of what happened? … I had a horrible headache. …How many times in the past month has that happened? …Those two times. …How long did it take you to sort of feel back to normal? I swear. It took me a minute. I got up. I got a glass of water. It took me about. I say two to three hours. …So how bad was that? Headache. Do you think there are any other symptoms? It was extremely. I never had. I had it like that.*

**Gold:** 4 (Extreme, dramatic physical reactivity, sustained arousal even after exposure has ended)
**GPT:** 3 (Pronounced, marked physical reactivity, sustained throughout exposure)
**Llama:** 2 (Clearly Present, physical reactivity clearly present, may be sustained if exposure continues)

*An example for Misaligned Reasoning.*

**Q:** the intensity of physical reactions in the past month.

**History:**
*… thinking about your work in the past month, how have you been doing? … It's a normal, consistent, um, it's a normal, consistent routine where I do the same thing, do the same thing every day.*

**Gold:** 40
**GPT:** NA (insufficient information)
**Llama:**40

*An example for Commonsense Reasoning.*

## CONCLUSIONS

- Create a new dataset comprising over 700 hours of clinician-administered PTSD interviews.
- Develop a novel and comprehensive pipeline to process the interview dataset.
  - Can be adapted to a broader range of diagnostic interviews.
- Develop assessment models that achieves promising results.