Supplementary Material

Supplementary

Effect of Few-Shot Sampling Strategies on PCL Score Estimation Performance

This section examines how different few-shot sampling strategies affect model performance in PTSD severity estimation using 70B-scale models. As shown in Table S6, we compare zero-shot prompting to two few-shot configurations: percentile-based sampling (selecting examples at the 25th, 50th, and 90th percentiles of the PCL score distribution) and range-based sampling (drawing examples from both the center and tail of the distribution). While few-shot prompting generally improves performance for base models like LLaMA-3.1-Base, its effect is less consistent for instruction-tuned models. For instance, LLaMA-3.1-Instruct shows no improvement, or even slight degradation, when few-shot exemplars are added. Across models, percentile sampling tends to slightly outperform range sampling, but gains are modest and vary by architecture. These results highlight that few-shot prompting is not uniformly beneficial and that the choice of example distribution is a sensitive factor in prompt design for clinical prediction tasks.

Comprehensive Performance Comparison Across Model Sizes and Prompting Strategies

Table S7 presents a full comparison of model performance across three model size categories (8B, 70B, and larger or undisclosed-scale models), evaluated under consistent prompting protocols. Each model is tested with zero-shot and 3-shot prompting, including standard, instruction-tuned, and chain-of-thought variants. We report both original and redistributed scores for Pearson correlation and MAE. This consolidated result table provides a reference point for understanding how architectural choices, prompting style, and inference strategies jointly shape LLM performance on PTSD severity estimation.

Optional Plug-In Components for Prompt Customization

Table S8 summarizes the optional plug-in components used to construct prompts across experimental configurations. These modular elements were toggled on or off to systematically assess their contribution to model performance in PTSD severity estimation. Each component provides a distinct type of clinical or contextual framing,

Table S6: **70B models**: We report Pearson correlation (higher is better) and mean absolute error (MAE, lower is better) under three prompting conditions: zero-shot, few-shot with percentile-based sampling (examples at the 25th, 50th, and 90th percentiles of the empirical PCL distribution), and few-shot with stratified sampling (two examples drawn from the central bulk of the distribution and one from its skewed tail).

Model Variant	Size	Prompt	Pearson ↑	MAE ↓
LLaMA-3.1-Base	70B	0-shot	0.147	31.79
		3-shot (range sampling)	0.345	16.73
		3-shot (percentile sampling)	**0.346**	15.49
LLaMA-3.1-Instruct	70B	0-shot	0.426	10.95
		3-shot (range sampling)	0.403	17.07
		3-shot (percentile sampling)	**0.430**	15.93
LLaMA-3.1-Instruct (TSBS)	70B	0-shot	0.412	11.59
		3-shot (range sampling)	**0.413**	13.33
		3-shot (percentile sampling)	0.396	12.58
LLaMA-3.1-Instruct w/ distr. info	70B	0-shot	0.407	9.62
		3-shot (range sampling)	0.375	11.37
		3-shot (percentile sampling)	**0.414**	13.01
DeepSeek-Distil-LLaMA	70B	0-shot	0.354	11.65
		3-shot (range sampling)	**0.374**	11.75
		3-shot (percentile sampling)	0.347	12.38

such as subscale definitions, interview questions, or study background (e.g., post-9/11 context). Other elements, like distributional information or PCL item references, offer structural guidance or calibration cues to shape model outputs. This modular design enabled targeted evaluations of which kinds of information most effectively guide large language models in generating accurate and clinically meaningful predictions.

Prompt Architecture for PTSD Severity Estimation Experiments

Figure S4 presents the full layout of the prompt used for the PTSD severity estimation experiments. The design consists of modular components that were either fixed or toggled depending on the experimental configuration. Core elements shared across all prompts include the task instructions, the structured scoring system, a clearly defined two-step procedure outlining the expected scoring process, as well as the expected output format. Optional plug-in components, shown in gray, include subscale definitions, interview questions and asking for references to parts of the text that motivated the model's decisions. These components were included or excluded in a controlled manner to assess their contribution to model performance.

Table S7: **Full model comparison across sizes.** Pearson correlation (↑) and MAE (↓) for PCL score estimation across prompting strategies. We report both original predictions and redistributed scores. Models are grouped by scale.

Model Variant	Size	Pearson ↑				MAE ↓			
		0-shot Original	3-shot Original	0-shot Redistr.	3-shot Redistr.	0-shot Original	3-shot Original	0-shot Redistr.	3-shot Redistr.
8B Models									
LLaMA-3.1-Base	8B	.178	.251	.206	.247	25.44	26.85	10.00	10.10
LLaMA-3.1-Instruct	8B	.316	.218	.326	.217	17.21	18.72	9.30	10.04
LLaMA-3.1-Instruct (TSBS)	8B	.322	.364	.328	**.365**	17.18	14.19	9.28	**8.86**
DeepSeek-Distil-LLaMA	8B	.208	.178	.221	.182	18.98	14.09	9.90	10.22
70B Models									
LLaMA-3.1-Base	70B	.187	.345	.191	.346	31.79	16.73	9.42	9.41
LLaMA-3.1-Instruct	70B	.426	.403	**.434**	.429	10.95	17.07	**8.25**	8.49
LLaMA-3.1-Instruct (TSBS)	70B	.412	.413	.421	.424	11.59	13.33	8.55	8.53
DeepSeek-Distil-LLaMA	70B	.354	.374	.364	.376	11.65	11.75	9.05	8.80
Larger Model Variants									
LLaMA-3.1-Instruct	405B	.426	.361	.437	.366	13.80	14.69	8.45	8.99
LLaMA-3.1-Instruct (TSBS)	405B	.429	.416	**.438**	.415	11.35	12.38	**8.27**	8.51
DeepSeek-R1	670B	.368	.319	.371	.322	11.45	19.58	8.93	9.99
4o-mini	N/A	.331	.295	.336	.307	13.66	23.96	9.23	9.43
4o-mini (TSBS)	N/A	.311	.294	.314	.303	14.15	20.01	9.33	9.49
o3-mini	N/A	.383	.344	.386	.349	8.81	10.74	8.85	9.11

Table S8: Overview of optional plug-in components used in the prompt design. These components can be toggled on or off across configurations to evaluate their effect on model performance.

Plug-in Component	Description
w/ Evidence	Prompts the model to explain the reasoning behind each subscale severity score prediction by referencing relevant excerpts from the transcript.
w/ Subscale Definitions	Short descriptions for each of the four PTSD subscales (re-experiencing, avoidance, dysphoria, hyperarousal) to guide the model's interpretation.
w/ Interview Questions	The set of questions originally posed to participants during self-recorded interviews, aimed at eliciting responses indicative of PTSD symptoms.
w/ Study Context	An expanded instruction indicating that the interview relates to post-9/11 lived experiences and clarifying the clinical framing.
w/ PCL Items	The original 17 PTSD Checklist (PCL) items used for clinical scoring [**?**], listed to help the model align its predictions to standard symptom indicators.
w/ Distributional Information	A summary of expected score distributions to steer the model away from extreme outputs without strong textual evidence.

Instructions: Your task is to conduct an in-depth analysis of a provided transcript, with the goal of assessing the severity of PTSD symptoms based on its 4 subscales. The transcript is derived from a patient's self-recorded answers to a set of questions, designed to assess various aspects of psychological well-being, including PTSD-related symptoms.

Questions Asked:

1. How are you? Can you elaborate?
2. How's the family? Can you elaborate?

   .
   .
   .

13. Over the past 5 years what are the three worst things that happened to you and your family? Can you elaborate?

PTSD Subscales:

- **Re-experiencing:** Intrusive thoughts, flashbacks, nightmares, or distress when reminded of the trauma.
- **Avoidance:** Efforts to avoid thoughts, feelings, or external reminders associated with the trauma.
- **Dysphoria:** Persistent negative thoughts, feelings of guilt, emotional numbness, or loss of interest in activities.
- **Hyperarousal:** Increased irritability, hypervigilance, being easily startled, or difficulty concentrating and sleeping.

Scoring System:

- 0 - Not at all severe: The symptom is not present or negligible.
- 1 - Low severity: The symptom is present but slightly impacts the individual.
- 2 - Moderate severity: The symptom is noticeably present and moderately impacts the individual.
- 3 - High severity: The symptom is significantly present and strongly impacts the individual.
- 4 - Extreme severity: The symptom is overwhelmingly present and has a profound impact.

Steps:

- **Step 1:** Detect the PTSD subfactors in the text *and provide an explanation on how each symptom was identified.* Provide a score between 0 and 4.
- **Step 2:** Present the results as a nested JSON with severity scores *and explanations referencing relevant spans or contextual inferences.*

```
Expected JSON Output:

{    "Re-experiencing": {
       "Reason": ["<Reason(s) from transcript>"] ,
       "Severity Score": <score>
   },
   "Avoidance": {
       "Reason": ["<Reason(s) from transcript>"] ,
       "Severity Score": <score>
   },
   "Dysphoria": {
       "Reason": ["<Reason(s) from transcript>"] ,
       "Severity Score": <score>
   },
   "Hyperarousal": {
       "Reason": ["<Reason(s) from transcript>"] ,
       "Severity Score": <score>
   } }
```

Text: *"I've been having a rough time lately. Some days are better than others, but overall, I feel like I'm constantly on edge. I still think about what happened all the time, not always directly, but little things set me off. Loud noises, crowded places. I try to keep busy..."*

Fig. S4: Full layout of the PTSD subscale evaluation prompt. Colored boxes represent core components shared across all configurations. Gray boxes are optional plug-ins that can be included or excluded to control the amount of context provided.

Additional Plug-In Components for Further Prompt Customization

Figure S5 illustrates three plug-in prompt components that were selectively incorporated in different experimental conditions. These elements either modify core instructions or introduce additional clinical context to guide the model's scoring behavior. Panel (a) shows an updated instruction variant that embeds explicit study context, framing the task around participants' experiences following the 9/11 World Trade Center attacks. Panel (b) introduces the full list of PTSD Checklist (PCL) items. Panel (c) presents distributional guidance about typical PCL score ranges, intended to calibrate model predictions by discouraging extreme scores unless strongly supported by the input. These components were activated in a controlled manner across configurations to investigate whether they yield more accurate severity estimates.

(a) Study Context [*Updated Instruction Component*]:
Your task is to conduct an in-depth analysis of a provided transcript, with the goal of assessing the severity of PTSD symptoms based on its 4 subscales. The text you will analyze is the transcription of the patient's self-recorded answers to a set of questions. The patients are questioned about their lives after the World Trade Center disaster that happened on 9/11. These questions were designed to assess various aspects of psychological well-being, particularly symptoms associated with PTSD. The goal is to assess the overall PTSD severity related to the World Trade Center disaster.

(b) PCL Items [*Additional Component*]:
The PCL score is calculated based on 17 specific items, each representing a symptom commonly associated with PTSD. The 17 items are:

1. Repeated, disturbing memories, thoughts, or images of a stressful experience from the past.
2. Repeated, disturbing dreams of a stressful experience from the past.

   .
   .
   .

17. Feeling jumpy or easily startled.

(c) Distributional Information [*Additional Component*]:
According to domain experts, most individuals typically score between 17 and 30. Scores between 30 and 50 occur in fewer cases, while above 50 are relatively rare and typically indicate more severe or clinically significant symptom expression. You should take this distribution into account when making your severity judgments. Use higher scores only when there is clear textual evidence of substantial psychological distress or impairment.

Fig. S5: Further plug-in prompt components that were explored in this study. These elements provide additional information and are activated selectively. They can either be additional components (*Additional Component)* or alternative variants of a core component (*Updated Component*).

Direct score prediction prompt

Instructions: Your task is to conduct an in-depth analysis of a provided transcript, with the goal of estimating the overall severity of post-traumatic stress disorder (PTSD) symptoms experienced by the individual. The text you will analyze is a transcription of the patient's spoken answers to a predefined set of self-recorded questions. The patients are questioned about their lives after the World Trade Center disaster that happened on 9/11. These questions were designed to assess various aspects of psychological well-being, particularly symptoms associated with PTSD. The goal is to assess the overall PTSD severity related to the World Trade Center disaster.

Scoring System: Based on the content of the transcript, predict a single scalar PTSD severity score in the range 17 to 85, where:
- 17 represents minimal or no PTSD-related symptoms. - 85 represents extreme PTSD symptom severity across multiple domains of functioning.
This score should directly estimate the patient's PCL score, a widely used self-report measure of PTSD symptom severity.

Steps: Carefully analyze the transcript, considering the emotional tone, content, and any references to trauma-related symptoms or functional impairments. Then assign a single integer score between 17 and 85 that best reflects the overall PTSD severity of the individual related to the World Trade Center disaster.

```
Expected JSON Output: Return your answer in the following structured JSON format:

{ "PTSD Severity Score": <score> }
```

Text: *"I've been having a rough time lately. Some days are better than others, but overall, I feel like I'm constantly on edge. I still think about what happened all the time, not always directly, but little things set me off. Loud noises, crowded places. I try to keep busy..."*

Fig. S6: Full layout of the PTSD direct score prediction prompt, which is simpler compared to the subscales prediction one shown in Figure S4.

Dataset Descriptive Statistics

Table S9: Descriptive statistics of our dataset, including demographic information and basic audio characteristics of the self-recorded clinical interviews.

Category	Metric	Analysis Set
Demographics	Number of Recordings	1437
	Age (min / mean / max)	38 / 58.09 / 90
	Gender Ratio (F:M)	7.4 : 92.6
Audio Stats	Avg. Duration (min)	7.50 ± 4.1
	Avg. Word Count	697.93 ± 530.44