Automating PTSD Diagnostics in Clinical Interviews: Leveraging Large Language Models for Trauma Assessments

Sichang Tu,[1] Abigail Powers,[1] Natalie Merrill,[1] Negar Fani,[1]
Sierra Carter,[2] Stephen Doogan,[3] Jinho D. Choi[1]
[1]Emory University, Atlanta, GA, USA
[2]Georgia State University, Atlanta, GA, USA
[3]Doogood Foundation, New York, NY, USA
{sichang.tu, abigail.d.powers, natalie.merrill, nfani,jinho.choi}@emory.edu
scarter66@gsu.edu, sdoogan@rlsciences.com

Abstract

We collect 411 clinician-administered diagnostic interviews and develop a novel approach to obtain high-quality data. Furthermore, we build a framework for automating PTSD diagnostic assessments, leveraging two state-of-the-art LLMs, GPT4 and Llama2, which can be expanded to broader mental health conditions. Our results show promising potentials of LLMs to assist clinicians in diagnostic decision-making and validation. To the best of our knowledge, this is the first computational model that fully automates PTSD assessment on structured clinical interviews.

1 Introduction

Mental health has become a crucial aspect of overall health and well-being. However, a significant treatment gap has been exacerbated by the critical shortage of mental health workforce — averaging 13 for per 100,000 people (World Health Organization, 2021). This scarcity hampers access to diagnostics and subsequent interventions.

The development of Large Language Models (LLMs) has provided innovative solutions to the mental health care challenges (Hua et al., 2024), including condition detection (Zhang et al., 2022), support and counselling (Ma et al., 2023), and clinical decision-making (Fu et al., 2023). However, there are notable limitations in the current research scope. Research primarily targets prevalent conditions such as stress (Lamichhane, 2023) and depression (Qin et al., 2023), with scant attention to less common disorders like Post-traumatic Stress Disorder (PTSD). Additionally, most studies leverage data from social media (Yang et al., 2023), Electronic Health Records (EHRs) (Irving et al., 2021), and clinical notes (Kshatriya et al., 2021). Though few works utilize computer-patient interview on self-administered questionnaires (Galatzer-Levy et al., 2023), there is a lack of work on automatic diagnosis using systematic clinician-administered diagnostic interviews.

In this work, we apply the state-of-the-art LLMs, GPT4 (OpenAI, 2023) and Llama2 (Touvron et al., 2023), to automate PTSD assessments in real-world clinician-administrated interviews. We introduce a novel approach to process diagnostic interview data (see Section 2), and build a versatile framework for benchmarking automatic PTSD diagnosis (see Section 3). This approach, potentially adaptable to broader mental health conditions, illustrates LLMs' capacity to expedite diagnostics, promising reduced clinician workload and minimal supervision requirements, thereby enhancing the efficiency of the diagnostic process.

2 Dataset

PTSD Interview Dataset 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 (Gluck et al., 2021). 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 (DSM-5; Weathers et al. (2018)). Every section comprises a set of questions, linked to variables that store pertinent values derived from the corresponding answers. Appendix A gives descriptions for all sections.

Data Processing Every video is converted into an MP3 audio file and *transcribed* by two auto-

Type	Count	Acc		RMSE		Bias		Recall	
		ChatGPT	Llama2	ChatGPT	Llama2	ChatGPT	Llama2	ChatGPT	Llama2
Scale	9,722	59.2	46.7	1.09	1.25	48.0	75.7	-	-
Scale_group	9,722	67.6	59.0	0.85	1.01	48.5	75.6	-	-
Category	4,258	77.2	63.6	-	-	-	-	-	-
Calculation	3,482	64.4	56.5	-	-	33.0	49.8	-	-
Notes	1,146	-	-	-	-	-	-	48.1	52.7
Rule_based	2,828	56.3	42.1	0.96	1.20	41.3	74.3	-	-

Table 1: Accuracy, Root Mean Square Error (RMSE), positive Bias Evaluation, and Recall achieved by both models for each variable type. Note that the results of Rule_based results derive from related Scale and Calculation variables. Scale_group is evaluated on grouped scales by clinical thresholds for the CAP section.

matic speech recognizers, Azure Speech-to-Text[1], and OpenAI Whisper (Radford et al., 2023) whose results are *aligned* by *align4d*[2] to produce a high-quality transcript. The transcript is *segmented* into multiple sections based on the relevant variables, and each variable is *paired* with its assessment result collected by the interview data manager REDCAP[3]. See Appendix B for detailed data statistics.

3 Experiments

We utilize 2 popular state-of-art large language models: Meta's *Llama-2-70b-chat-hf*, the largest Llama2 model, and OpenAI's *gpt-4-1106-preview* (ChatGPT), the latest GPT-4 Turbo model, for evaluating interview data.

Prompt Engineering To streamline the prompting generation process, we develop templates based on variable types for universal application. Each template includes one or more adaptable patterns, specifically crafted based on the instructions of the interview question set (See Appendix C). We also investigate zero-shot and few-shot settings on ChatGPT and Llama2 models, assessing performance across variable types. ChatGPT performs better in few-shot settings, whereas Llama2 excels in zero-shot scenarios. Consequently, we employ few-shot prompting for ChatGPT and zero-shot prompting for Llama2.

Parameters For ChatGPT, we set the temperature to 0 for deterministic outputs and use exclusive parameters like response_format as "json_object" and seed for consistent responses. Llama2 experiments with temperature at 0.3, top_p at 0.9, and repetition_penalty at 1 show improved performance.

Metrics We evaluate model performance using tailored metrics for different variable types. Accuracy measures effectiveness across most variables, where Recall assesses information coverage for Notes variables. For Scale variables, Root Mean Square Error (RMSE) quantifies prediction errors, and Bias Evaluation identifies directional prediction biases, offering a comprehensive view of model accuracy and reliability.

4 Results

Table 1 shows that ChatGPT outperforms Llama2 by an average of 10% across most metrics, except for Note variables where Llama2 yields higher recall. Specifically, ChatGPT achieves higher accuracy and lower RMSE in Scale variables, indicating closer alignment with gold data but tends to predict more conservatively. In contrast, Llama2 exhibits a tendency to overestimate, reflected in its higher positive bias. ChatGPT also leads in the accuracy of Category and Calculation variables. Rule variables maintain the similar trend in all metrics to Scale variables, as their outcomes are calculated based on the predictions of Scale and Calculation variables.

5 Conclusion

In this work, we introduce a novel pipeline for structured clinician-administered interview data, and establish the framework for automating PTSD diagnostics. This is the first study have evaluated the effectiveness of LLMs in PTSD psychiatric screening, offering new possibilities in mental health diagnostics.

[1] https://azure.microsoft.com/en-us/products/ai-services/speech-to-text

[2] https://github.com/emorynlp/align4d

[3] https://www.project-redcap.org

References

Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, Juan Zhang, and Bing Xiang Yang. 2023. Enhancing Psychological Counseling with Large Language Model: A Multifaceted Decision-Support System for Non-Professionals.

Isaac R. Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The Capability of Large Language Models to Measure Psychiatric Functioning.

Rachel L. Gluck, Georgina E. Hartzell, Hayley D. Dixon, Vasiliki Michopoulos, Abigail Powers, Jennifer S. Stevens, Negar Fani, Sierra Carter, Ann C. Schwartz, Tanja Jovanovic, Kerry J. Ressler, Bekh Bradley, and Charles F. Gillespie. 2021. Trauma exposure and stress-related disorders in a large, urban, predominantly african-american, female sample. *Archives of Women's Mental Health*, 24(6):893–901.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yi-han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, and Andrew Beam. 2024. Large Language Models in Mental Health Care: a Scoping Review.

Jessica Irving, Rashmi Patel, Dominic Oliver, Craig Colling, Megan Pritchard, Matthew Broadbent, Helen Baldwin, Daniel Stahl, Robert Stewart, and Paolo Fusar-Poli. 2021. Using natural language processing on electronic health records to enhance detection and prediction of psychosis risk. *Schizophrenia bulletin*, 47(2):405–414.

Bhavani Singh Agnikula Kshatriya, Nicolas A Nunez, Manuel Gardea Resendez, Euijung Ryu, Brandon J Coombes, Sunyang Fu, Mark A Frye, Joanna M Biernacka, and Yanshan Wang. 2021. Neural Language Models with Distant Supervision to Identify Major Depressive Disorder from Clinical Notes.

Bishal Lamichhane. 2023. Evaluation of Chatgpt for Nlp-based Mental Health Applications.

Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1105. American Medical Informatics Association.

OpenAI. 2023. Gpt-4 Technical Report.

Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, pages 28492–28518.

David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, Geoffrey C Dunbar, et al. 1998. The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *Journal of clinical psychiatry*, 59(20):22–33.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Frank W. Weathers, Michelle J. Bovin, Daniel J. Lee, Denise M. Sloan, Paula P. Schnurr, Danny G. Kaloupek, Terence M Keane, and Brian P. Marx. 2018. The Clinician-Administered PTSD Scale for DSM-5 (CAPS-5): Development and initial psychometric evaluation in military veterans. *Psychological Assessment*, 30(3):383–395.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

World Health Organization. 2021. *Mental health atlas 2020*. World Health Organization.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards Interpretable Mental Health Analysis with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine*, 5(1).

Section	Questions	Variables	Example Question	Example Variable
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 or 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 trauma1_distress

Table 2: Statistics and examples for each of the four sections employed in this study.

A Section Descriptions

There are 10 different sets of assessment question sets. Out of these 10 sets, 4 core question sets are applied to the majority of the participants, while the remaining 6 are optional and are utilized as needed. Table 2 shows statistics and examples for each of the 4 core sections.

LBI It assesses the participant's functioning over the past month, addressing topics such as daily life, work, relationships with friends and family, and overall life satisfaction.

THH It covers the participant's treatment/health history, including past physical and mental conditions as well as treatments received, such as medication and therapeutic services.

CRA It assesses whether the participant has been exposed to (threatened) death, serious injury, or sexual violence, with a focus on potential traumatic experiences the participant might have endured.

CAP It centers on issues the participant may have encountered due to traumatic events, including distress, avoidance of trauma-related stimuli, negative thoughts and feelings, and trauma-related arousal.

MINI The Mini International Neuropsychiatric Interview (MINI) is a brief, structured diagnostic interview for diagnosing 17 major psychiatric disorders (Sheehan et al., 1998). We adopt 6 modules from MINI to assess conditions such as Major Depressive Episode, Mania & Hypomania, PTSD (past incidents), Psychosis Symptoms, Substance Use Disorder, and Alcohol Use Disorder.

B Data Statistics

We collect a total of 456 diagnostic interview audios, recorded using online conferencing software such as Zoom and Microsoft Teams. Each interview lasts 1.5 hours on average, involving the participant and 1-2 interviewers. Azure Speech-to-Text and OpenAI Whisper are employed for Automatic Speech Recognition (ASR). Whisper demonstrates better performance in handling noisy environments and numbers (e.g., dates, times, ages) than Azure that often misses or inaccurately transcribes them. Despite its superior performance in ASR, however, Whisper lacks the capability for speaker diarization, a feature found in the other tools. Thus, both Azure and Whisper are run on all converted audios and their results are combined to obtain the best possible outcomes. While state-of-the-art, Whisper tends to generate irrelevant or repetitive sequences when prolonged silences occur, rendering unusable transcripts. To address this issue, we apply silence removal and noise cancellation to these audio files before transcription, successfully recovering the majority of them. As a result, a total of 393 interviews are transcribed for our experiments.

	Audios	Hours	Turns	Tokens
Original	411	702	233,002	6,035,027
Transcribe	393	651	180,347	5,499,662
Evaluation	322	512	142,824	4,335,977

Table 3: Statistics of interviews and transcripts in the original data and the final data used for our experiments.

Var Type	Variables					Count
	LBI	THH	CRA	CAP	Total	
Scale	7	1	0	40	48	9,722
Category	4	9	15	3	31	4,258
Calculation	2	0	1	24	27	3,482
Notes	1	10	3	0	14	1,146
Rule-based	1	0	1	25	27	2,828

Table 4: Statistics of paired variables for each section.

We map the segmented section data with vari-

Var	Template
S & Cate	Imagine you are a professional clinician. Based on the patient's interview history, please determine {keywords} that the patient {symptom}. Return the answer as a JSON object with "reason" and "answer" as the keys. The "reason" should provide a brief justification or explanation for the answer. The "answer" should be in the range {range}.{attributes}
Cal	Imagine you are a professional clinician. Based on the patient's interview history, please calculate {keywords} that the patient have {symptom}. Return the answer as a JSON object with "reason" and "answer" as the keys. The "reason" should provide a brief justification or explanation for the answer. The "answer" should be {type}.
Notes	Imagine you are a professional clinician. 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 answer as a JSON object with "reason" and "answer" as the keys. 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.

Table 5: System message templates for **S**cale variable, **Cate**gory variable, **Cal**culation variable, and **Notes** variable.

able score data exported from the REDCap system. The variables within this system are categorized into five distinct types, each based on their value types and requiring different prediction and evaluation methods. Each question set contains multiple categories of variables, and they should be predicted and evaluated differently. Table 4 gives the statistics of the paired data.

`Scale Variable` is assessed on an ordinal scale with ratings indicating the intensity, severity or likeness. Although the values of `Category Variable` are often ordinal numbers, they represent either binary choices (e.g. yes or no) or distinct class labels. `Calculation Variable` includes all variables that ask about duration, frequencies, and ages, which may require mathematical calculations. `Notes Variable` is summarized text that is manually documented by clinicians during the interview. Derived from other variables, `Rule-based Variable` is calculated on predefined rules.

C Prompt Templates

As shown in Table 5, each template includes one or more patterns that are dynamically replaced with specific patterns to generate the final prompts. These patterns are hand-crafted according to interview question set instructions. For instance, in *Scale Variable* template, the "keywords" patterns like "how severe" and the "symptom" pattern such as "have unwanted dreams in the past month" can be adapted. Once we replace the "keywords" to "which of the following categories best describes" and "symptom" to "usual employment status", the prompt is applicable for *Category Variable*. To better instruct the model, we incorporate details such as answer "range" for *Scale* and *Category Variable* and value "type" for *Calculation Variable* to restrict the answer returned by the model. We also provide additional brief explanation to each level in the range of *Scale Variable*. The "attribute" pattern is exclusive for *Scale Variable*, directing the model to return a particular score under certain conditions.

Different from other variables, *Notes Variable* are evaluated against the gold summarized text data. The prediction encompasses multiple tasks, including information categorization, information extraction and multiple times of binary classification. Since such a complicated task might degrade the performance, we decide to break down the process, inspired by the Chain-of-Thought (CoT) Prompting technique (Wei et al., 2023). We first ask the model to generate a list of slots for each *Notes Variable* with a batch of gold summary data as the user input. Although these slots may vary in naming, they largely overlap. Hence, we pass all slots to the model and merge them into groups. The final grouped slots we adopt cover over 95% of slots of the initial generation for each variable, which balances the brevity of the list with comprehensive coverage. With these predefined slots, we format both gold summarized text and the corresponding interview history into the same structure. This structure allows the model to verify the presence of each slot in the interview data against the gold formatted data.