Using LLMs to Aid Annotation and Collection of Clinically-Enriched Data in Bipolar Disorder and Schizophrenia

Ankit Aich[1, 5, 6], **Avery Quynh**[2], **Pamela Osseyi**[5], **Amy Pinkham**[3], **Philip Harvey**[4], **Brenda Curtis**[5], **Colin Depp**[2], **Natalie Parde** [1],
[1]Department of Computer Science, University of Illinois Chicago,
[2]University of California, San Diego,
[3]University of Texas Dallas, [4]University of Miami,
[5] National Institute on Drug Abuse, National Institutes of Health,
[6]School of Engineering and Applied Science, University of Pennsylvania

Abstract

Natural Language Processing (NLP) in mental health has largely focused on social media data or classification problems, often shifting focus from high caseloads or domain-specific needs of real-world practitioners. This study utilizes a dataset of 644 participants, including those with Bipolar Disorder, Schizophrenia, and Healthy Controls, who completed tasks from a standardized mental health instrument. Clinical annotators were used to label this dataset on five clinical variables. Expert annotations across five clinical variables demonstrated that contemporary language models, particularly smaller, fine-tuned models, can enhance data collection and annotation with greater accuracy and trust than larger commercial models. We show that these models can effectively capture nuanced clinical variables, offering a powerful tool for advancing mental health research. We also show that for clinically advanced tasks such as domain-specific annotation LLMs provide wrong labels as compared to a fine-tuned smaller model.

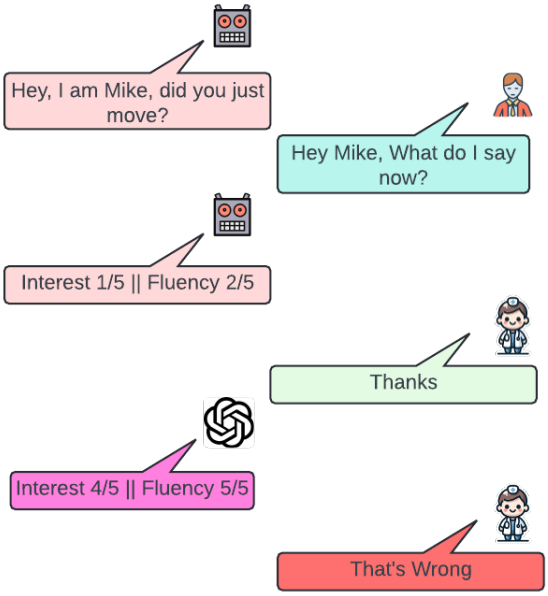

Figure 1: Our method creates a fine-tuned model. This model is able to directly interact with recruited participants to help them undertake established mental health instruments through turn-based tasks. It can annotate for clinical variables with low error. We see that commercial LLMs like GPT-4 / GPT-4o cannot annotate when it comes to clinical variables which are niche to a domain.

1 Introduction

The inherent complexity of mental health data presents significant challenges, even as the availability of AI systems designed to aid in its understanding and categorization continues to grow (Lee et al., 2021). AI-based systems have increasingly leveraged social media as a data source in the realm of mental healthcare, leading to the development of pre-trained models like MentalBERT (Ji et al., 2022) and initiatives to classify and detect various mental health phenomena, such as schizophrenia (Liu et al., 2022), disease progression (Birnbaum et al., 2019), depression (Kang et al., 2016), and stress (Winata et al., 2018).

In addition to the ethical issues surrounding the use of social media for clinical diagnoses, numerous other challenges persist. These include participant bias (Palacios-Ariza et al., 2023), issues with generalizability (Mitchell et al., 2015), and an overreliance on self-disclosure or non-clinical labels (Mitchell et al., 2015; Coppersmith et al., 2014).

Psychiatric disorders such as schizophrenia and bipolar disorder are often characterized by language deficiencies (Merrill et al., 2017). Individuals with these conditions may exhibit disorganized language comprehension and speech patterns (Kuperberg, 2010). Consequently, text or speech-based mental health instruments can be employed to assess individuals with medically validated diagnoses, thereby elucidating the effects of psychiatric disorders.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 181–192

Previous efforts to apply AI in the context of schizophrenia and bipolar disorder have predominantly focused on automated diagnoses using smaller datasets (Sadeghi et al., 2021). These classification endeavors have encountered multiple challenges. For instance, social media data often results in non-clinical labels (Ernala et al., 2019), while the classification of clinical data is complicated by small datasets (Sadeghi et al., 2021), underutilization of records (Montazeri et al., 2022), and attempts to apply AI to multiple psychiatric disorders simultaneously (Chandran et al., 2019).

Moreover, the scarcity of robust data sources for mental health care and AI remains a significant barrier, as noted by Harrigian et al. (2021). The reliability of social media labels is further undermined over time due to evolving subjective annotation metrics (Harrigian and Dredze, 2022). To enhance the application of AI and language models in schizophrenia and bipolar disorder research, we propose a novel approach. This approach involves testing the efficacy of AI models in the context of data collection and annotation.

Our study starts with a dataset comprising 644 participants with established medical histories of schizo-affective disorder or schizophrenia (SZ), bipolar disorder (BD), or who are healthy controls (HC). These participants undergo a mental health instrument involving interviews conducted by expert clinicians (Patterson et al., 2001). We engaged two expert clinicians to annotate transcribed speech samples across five clinical variables. Importantly, we do not conduct automated diagnoses nor suggest that language models should be used for diagnostic purposes. Instead, we demonstrate how modern language models can assist in data collection and annotation.

The contributions of this paper are as follows:

- Extending a real-world dataset with expert clinical annotation, focusing on the language and speech deficiencies of individuals with bipolar disorder and schizophrenia.
- Creating a model that assists clinicians in maintaining dialogue with recruited participants for data collection purposes.
- Creating another model replicating clinical annotation of domain-specific variables with low error.
- Demonstrating that our models achieve low error rates and higher accuracy compared to commercial language models like GPT-4.

2 Data Collection and Labeling

We start by using the dataset introduced by Aich et al. (2022) in 2022. The data consists of transcribed texts from interviews with 644 participants. In the initial dataset, the authors recruited participants from three categories: participants with schizophrenia, participants with bipolar disorder, and healthy control groups. The diagnoses for subjects are all based on the DSM-V. The participants were in two simulated clinical tasks with expert clinicians to build the dataset. For task descriptions, please refer to appendix A.

We present a clinical annotation task to expand the dataset.

2.1 Clinical Annotation of Data

We collect clinical scores for our SSPA data. The SSPA instrument variables (Mausbach et al., 2008) are defined below. Annotators adhering to these definitions were found to have near-perfect agreement $\{\kappa \geq 0.85\}$ when labeling the presence of these variables (Patterson et al., 2001):

- **Interest/Disinterest:** Subjects with a relevant mental health condition show low engagement in SSPA tasks since brain functions are impaired.
- **Fluency:** Subjects with higher fluency use fewer filler words such as *umm*, *you know*, or *sooo*, and/or fewer long pauses during SSPA tasks.
- **Clarity:** Subjects with greater communication clarity exhibit stronger coherence in speech, both in how things were said and what was said. In lay terms, this variable describes how well subjects can get their point across.
- **Focus:** Subjects with greater focus can more solely concentrate on the task given to them without veering from their course. This variable also describes the subject's ability to focus on the interviewer and the current and overall task objectives.
- **Social Appropriateness:** Subjects with greater social appropriateness scores fare better socially with respect to the scene. They

react more appropriately to interview cues and are able to maintain increased composure during tasks.

These five SSPA scores are based on participants' interactions with the clinicians. Each of these scores is annotated for a subject in each scene. The scores are then averaged across the scene for the subject. A subject's total SSPA score is the average of their two scene scores. Scoring is performed manually by experts, achieving a high inter-class coefficient. As shown in prior work (Patterson et al., 2001), subjects' SSPA scores are significantly correlated with the presence or absence of schizophrenia/schizoaffective disorder ($p < 0.01$).[1]

For annotation and collection, there were two expert annotators. These were practicing clinicians and researchers in psychiatry. Each annotator reviews the entire transcript and labels all five scores. Gold standard labels are adjudicated by discussion among clinical experts. The SSPA is a well-established standardized test with the scoring metrics clearly defined. Cohen's Kappa κ for all clinical scores was $\kappa \geq 0.85$. For our work, we consider the final adjudicated gold standard labels.

3 Methods - Interview Sequence Generation

3.1 Context-Aware Interviewer

Our first specialized objective was to design a proof-of-concept context-aware interviewer to facilitate SSPA sessions. Currently the SSPA is administered by human clinicians with heavy case-loads. The US mental healthcare system is already heavily over burdened with a very low number of clinicians to a high number of patients (Coombs et al., 2021), potentially leading to mistakes and reduced efficiency. Having a trustworthy and viable agent can help alleviate some of this. To administer the SSPA in a language modeling setting understanding of context is important. Each response from an interviewer depends not only on the previous turn, but the entire dialogue history to that point, i.e. the entire context window of that string. As described previously, our SSPA data is represented as two sets of dialogues (lists of n utterances), one of which belongs to the patient P and the other of which belongs to the interviewer I: $P = \{P_0, P_1, ..., P_n\}$ and $I = \{I_0, I_1, ..., I_n\}$. Both are stored with associated timestamps indicating when utterances begin.

In a real world setting, it is expected that an interviewer has facilitated many interviews before, across people with bipolar disorder and schizophrenia as well as people with neither condition. It is also expected that in each complete dialogue turn $\{P_i, I_i\}$, the Interviewer response I_i is not only a response to the dialogue P_i but to the set of dialogues $\{P_0, I_0, P_1, I_1, ..., I_{i-1}, P_i\}$. The intuition is thus that the interviewer is responding not only to what was just uttered by the patient, but in a way that is suitable with the entire conversation so far, including all patient and interviewer utterances up to the most recent patient utterance.

3.2 Task Setup for Interview Experiment

In this section we describe our setup for the supervised fine-tuning (SFT) experiment. We model this task as a sequence to sequence problem. Our model is trained to generate an appropriate sequence of dialogue in response to dialogue sequences it has seen in such a way that it is aligned with that generated by a real-world interviewer. We train on 75% of all BD, HC, and SZ dialogues across both scenes. The input and outputs for the encoder-decoder forward pass are:

$$I \rightarrow Out = \begin{cases} P_0 \rightarrow I_0, & \text{if } n = 0 \\ P_0, I_0, ..., I_{i-1}, P_i \rightarrow I_i, & n = i \end{cases}$$

The equation above again emphasizes that to create input-output pairs we consider the dialogue history in addition to the most recent utterance. If we are at index 0 of a conversation, the interviewer's response is based directly on the the patient's utterance, but otherwise the interviewer response is based on the entire dialogue history between the patient and interviewer, until the *i-1*th interviewer utterance and the patient utterance P_i.

A schematic diagram for the SSPA language modeling process is shown in Figure 2. The model is fine-tuned until the loss drops from 1.64 initially to 0.1 after 15000 checkpoints and then results are calculated. To initialize training we provide the following source prefix:

```
You are an intelligent interviewer see the examples provided and learn to interview a new patient
```

[1] Results are from a t-test taken comparing SSPA scores for schizophrenia and control group patients.

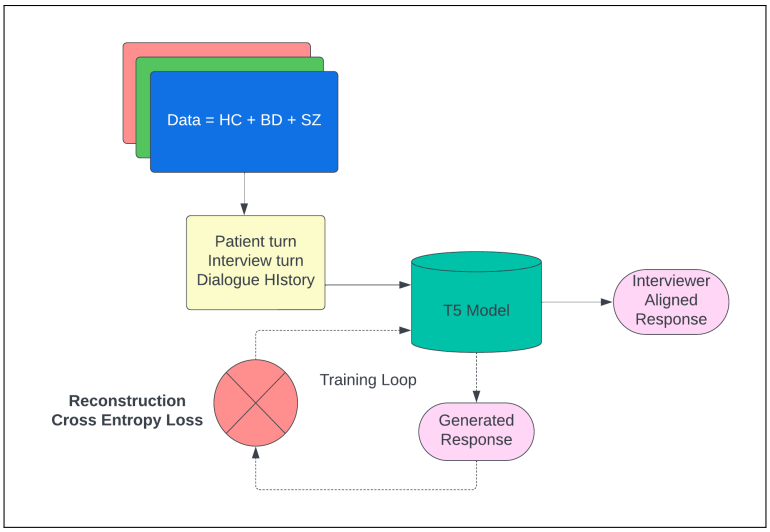

Figure 2: Interview model turns and dialogue history to calculate reconstruction loss and generate well aligned sequences towards the SSPA

We selected this prefix after experimenting with simpler versions (e.g., *"Interview a patient"* and *"Talk to a patient based on examples"*) and finding that the more complex final prefix was necessary to produce results well-aligned with reference interviews. This process does not involve in-context learning (ICL), prompt engineering, or tuning; it is a manually constructed prefix. Current literature suggests that better prefix descriptors, followed by improved training, yield superior results (Xue et al., 2022). Standard hyper-parameters were maintained at default values, and training was conducted on T4 GPUs.

The fine-tuned model was tested individually on all scenes and classes to simulate real-world interview conditions, where the interviewer focuses on a single scene and person, independent of prior interview training. To evaluate the quality of generated output, we computed syntactic similarity using ROUGE (Lin, 2004) scores, semantic similarity using cosine similarity, and alignment with human dialogue using BERTScore (Zhang et al., 2020).

3.3 Results: Generated Interview Quality

In Table 1 we present the results of the interview SFT experiment. To compute semantic similarity scores, we first encoded both the generated utterance and the corresponding gold expected utterance as word embeddings from the DeBERTA model owing to the model's increased ability to align with human speech (Zhang et al., 2020; He et al., 2021), and then calculated the cosine similarity between those embeddings. To compute syntactic similarity, we use ROUGE-1 to find overall unigram overlap and ROUGE-L to find the longest common subsequence overlap. We report precision, recall, and F1 scores for these two metrics using the ROUGE-SCORE package from the Python library.[2] We use the BERTScore (Zhang et al., 2020) metric directly, using a deberta model to vectorize the inputs to the metric generator (Zhang et al., 2020),[3] and report the precision, recall, and F1-score. According to the authors of the original paper, this model (deberta) offers the best understanding of the closeness of generated text to human intent. For all semantic metrics we use deberta as our choice of model since it has been consistently shown to outperform other encoder based popular choices like BERT or RoBERTa (He et al., 2021).

BERTScores, designed to capture intent and semantic similarity, are almost double the corresponding ROUGE scores for the same scenes. Recent studies have shown (Zhang et al., 2020; Hanna and Bojar, 2021) that BERTScore has two important properties. Firstly, it correlates with other summarization and similarity metrics (e.g., cosine similarity or BLEU score). Secondly, when a task becomes harder such as in our case, BERTScore accuracy peaks around 80% (Hanna and Bojar, 2021). Considering that our BERTScores for our task are close to 70% we can conclude that our model works at a high performance level. A better cosine simi-

[2] https://pypi.org/project/rouge-score/

[3] BERTScore needs users to specify which model to use to calculate metrics between two given strings. We use deberta for the same reasons cited earlier; i.e., studies have found it to generate text that more closely matches human speech.

Class × Scene	Semantic	Syntactic Similarity						Human Alignment		
	Cosine	ROUGE-1			ROUGE-L			BERTScore		
		P	R	F1	P	R	F1	P	R	F1
BD Scene_1	0.652	0.381	0.380	0.360	0.363	0.370	0.340	0.66	0.66	0.66
BD Scene_2	0.623	0.361	0.346	0.334	0.344	0.336	0.317	0.61	0.61	0.61
SZ Scene_1	0.634	0.331	0.316	0.301	0.328	0.314	0.300	0.63	0.64	0.63
SZ Scene_2	0.613	0.371	0.362	0.346	0.360	0.352	0.340	0.62	0.63	0.61
HC Scene_1	0.670	0.390	0.390	0.360	0.380	0.390	0.360	0.67	0.68	0.67
HC Scene_2	0.643	0.402	0.392	0.380	0.390	0.380	0.370	0.64	0.64	0.63

Table 1: Interview SFT Results. *P*=precision, and *R*=recall.

larity represents closeness in the embedding space of the vectors, whereas a good BERTScore tells us that the outputs are aligned with the reference sample.

However, even with a well-performing model, our ROUGE score is quite low. Some of this may be attributed to hallucinatory effects. For example, we observe that in one case while the interviewer in the original script says, e.g., "My name is INTERVIEWER," our model generates, e.g., "My name is NAME"—that is, a hallucinated name that was never previously mentioned in the dialogue. Thus, although this is structurally aligned with the reference, it differs in a key way that is best captured by ROUGE.

Another reason why our model exhibited lower syntactic than semantic performance may lie in disfluency. In our reference dialogues the interviewers often pause using filler words like *uh*, *uhh*, *okay*, or *mmhmm* to give the patients more time to speak. While our model thematically aligns decently well with these statements, its exact filler word matches are quite low. For example, we observe that the model also pauses but uses different filler words, or longer sequences of filler words, negatively affecting our ROUGE metric. However, throughout our observations, we can see that the model seems to understand the SSPA expectation of the interviewer role, even though we do not specify this in our SFT setup explicitly.

We qualitatively observe that the model appears capable at staying on-topic for the scene-specific task (e.g., generating content like *"Of course I will try to send someone over the fix the leak."*). It is interesting to observe that the model can discern the underlying task over long periods of training. Even without telling the model explicitly what the SSPA task involves, we can see that the model understands that a leaky pipe is at its core. This may suggest that LLMs are well-suited for tasks with better data and longer training (Min et al., 2022; Brown et al., 2020). While our alignment-based scores are not exceptionally high, this is still a strong starting benchmark for a nuanced task (Hanna and Bojar, 2021). The model captures close to 70 points of alignment with the intent of the actual interviewer. In the next phase we use this to generate annotator scores using another model to further progress the autonomous pipeline.

4 Methods - Annotation Generation

We also frame our SSPA score prediction task as a sequence to sequence task. Rather than simply predicting a sequence of scores, we also predict the label for which the score is being generated. In Section §2.1 we discussed what the five clinical variables are and how the scores are collected. In this score prediction task, the model learns to predict the score (SSPA clinical value) and the corresponding label. Therefore, the model predicts a sequence $Interest = XX, Fluency = YY$ rather than a simple distribution $4, 5, 3...$ This increases complexity, but helps us evaluate and walk towards a more explainable model. For this setting we use the interview dialogues from our source dataset and a `t5-base` LLM. We use the following prompt to generate scores:

```
You are an intelligent annotator
see the examples provided and
generate scores for each variable
```

The source prefix selection and other model parameters are kept the same as in the interview generation task described earlier in this paper. The model trains for 10000 checkpoints and the validation loss goes from 0.8 to 0.02 in our best performing

model checkpoint. We calculate this reconstruction loss between the variable labels and scores that are annotated by our clinicians and the ones that are generated by our model. A standard cross-entropy loss function is used to find the loss. The model is trained on 75% of the data and validated on 5% of the data, with the remainder held out for testing.

4.1 Results: SSPA Score Prediction

The results of the SSPA score prediction model are presented in Table 2. The values represent the root mean squared error (RMSE) between the original annotated labels $Y = \{S_1, S_2, \ldots, S_n\}$ and the predicted labels $Y' = \{S'_1, S'_2, \ldots, S'_n\}$, where $Y \in \{\text{Interest}, \text{Fluency}, \text{Clarity}, \text{Focus}, \text{Social}\}$. The results indicate generally low error, with improved predictive performance in Scene 2 compared to Scene 1.

The model exhibits superior performance for the variables *Social* and *Focus*, which is anticipated as the SSPA predominantly evaluates social skills, and *Social* encapsulates social appropriateness. The variables *Focus*, *Clarity*, and *Fluency* are linguistically dependent, with the model performing best in this order, and the least effective for *Interest*. The higher RMSE for *Interest* can be attributed to its reliance on non-verbal cues such as body language, which are absent from our transcripts.

Overall, this standalone model demonstrates effective prediction capabilities. In the subsequent section, we illustrate the adaptation of our previous model from §3.1 into a chained pipeline, enabling SSPA interview transcripts to be scored with minimal RMSE differences compared to the standalone model.

5 Chained Model

So far in this paper we have created two standalone models: one in §3.1 that can learn from interviewers to appropriately interact with patients to facilitate the SSPA task, and the other in §4 that reads patient-interviewer transcripts and generates a sequence of SSPA scores for a patient. In this section we experiment with combining them. The primary motivation for this lies in anticipated real-world need, moving towards a seamless support tool for busy clinicians who may otherwise need to administer and score the SSPA manually. We create a chained model that (1) converses with the patient, and (2) predicts SSPA scores from the encounter.

We predict scores for dialogues that our model generated in §3.1. The input consists of the entire dialogue between the patient P and generated interviewer dialogues I_{gen}, forming the sequence $\{P_0, I_0, P_1, I_1, ...P_n, I_n\}$, where an interviewer dialogue $I_k \in \{I_{gen}\}$ acts as input and the model returns a sequence of five integer-valued scores, $\{S_1, S_2, .., S_5\}$, that quantify the SSPA variables defined in §2.1 (Interest, Fluency, Clarity, Focus, and Social).

5.1 Results: Chained Model

We present the results of the experiment in Table 3. The scores reported are the RMSE between the expected SSPA scores and the generated SSPA scores predicted for LLM-facilitated SSPA transcripts. Our first observation is the acute closeness to the stand-alone model scores (recall Table 2). This shows that even when LLM-based assistants are adapted in a chained end-to-end fashion, the results are similar to those observed using standalone models.

When we compare the difference between errors for Tables 2 and 3 we can see that the differences are quite low at both a variable level and a *class X scene* level. We can see in Tables 5 and 4 that on a per scene or per variable basis the differences are quite low with no significant difference [4]

6 Comparison with GPT Models

To compare the performance of our model against a large model like GPT, below we provide a baseline comparison between GPT-4, GPT-4o, and our method in replicating annotation tasks as detailed in §2.1. To get these labels we show GPT-4 the same de-identified data along with definitions of the clinical variables and ask it to label the data along these five categories. The results, presented in Table 6, illustrate the mean errors per class and scene, with statistical significance validated using the Wilcoxon signed-rank test. We find that GPT models show a high degree of error in comparison to our method when annotating clinical scores. This shows that a small fine-tuned model can outperform a large model like GPT with appropriate fine tuning.

Our experiments reveal two significant trends in our interview replication model: an intrinsic comprehension of tasks and the generation of unrelated

[4] A t-test between the RMSE scores per case (scene and class) and per variable shows the differences between score distributions for the standalone and chained models are not statistically significant ($p < 0.05$).

Class and Scene	RMSE					
	Interest	**Fluency**	**Clarity**	**Focus**	**Social**	**Avg. RMSE/Case**
BD Scene_1	1.36	1.10	1.04	0.97	1.06	1.10
BD Scene_2	1.09	1.11	1.14	1.15	1.12	1.12
SZ Scene_1	1.27	1.27	1.28	1.19	1.30	1.26
SZ Scene_2	1.22	1.10	1.13	1.10	1.07	1.12
HC Scene_1	1.28	1.36	1.35	1.33	1.33	1.33
HC Scene_2	0.84	0.78	0.68	0.84	0.68	0.76
Avg. RMSE/Var	1.17	1.12	1.10	1.09	1.09	N/A

Table 2: RMSE scores for standalone score prediction model, using original dataset. Avg-RMSE/Case represents the mean RMSE across a diagnostic group and scene. Avg-RMSE/Var represents the mean RMSE for that SSPA variable of the column.

Class and Scene	RMSE					
	Interest	**Fluency**	**Clarity**	**Focus**	**Social**	**Mean/Case**
BD Scene_1	1.28	1.12	1.07	0.97	1.06	1.10
BD Scene_2	1.39	1.11	1.14	1.18	1.10	1.18
SZ Scene_1	1.37	1.33	1.27	1.20	1.30	1.29
SZ Scene_2	1.33	1.13	1.12	1.15	1.10	1.16
HC Scene_1	1.33	1.37	1.27	1.30	1.28	1.31
HC Scene_2	0.83	0.78	0.75	0.92	0.75	0.80
Avg. RMSE/Var	1.25	1.14	1.10	1.12	1.09	N/A

Table 3: RMSE scores for the chained score prediction model. Interview sequences come from the generative model described in §3.1. Mean/Case represents the mean RMSE across a diagnostic group and scene. Avg-RMSE/Var represents the mean RMSE for that SSPA variable of the column.

information. Even without explicit task instructions, a well-constructed fine-tuning loop allows a smaller model to intuitively understand tasks, evidenced by the model's ability to identify tasks from indirect references. Despite the tendency of the model to hallucinate information such as names and dates, which typically impedes performance on tasks necessitating factual precision, our findings indicate that these hallucinations do not compromise task completion. For our annotation task, we maintained a sequence-to-sequence setup for predicting scores and variable labels, observing low error rates and consistent performance across both stand-alone and chained model setups.

7 Conclusion

This paper focused on an alternate purpose of LLMs in mental healthcare. Instead of classification or diagnostic problems we focus on a collaborative-LLM setup. We show that for real world clinical tasks, often involving complicated and nuanced variables, smaller and focused fine-tuning can help with data collection and annotation with relatively low error. We also show that such models can be chained together to create reliable and robust end-to-end data collection and annotation pipelines. We showed that modern LLMs such as GPT-4 or GPT-4o do not perform at the same level as a fine-tuned model on clinically nuanced tasks.

In mental health settings the expertise that clinicians bring cannot be replaced by LLM technology. Rather a collaborative approach where locally trained LLMs can learn from clinical labeling behavior without compromising data to external servers is a better way forward. Our findings indicate that language models can significantly assist clinicians in scaling data collection and labeling with high reliability, as evidenced by low error rates and high similarity scores. We anticipate that the clinical community will find our models ready for practical implementation, and our methods both translatable and adaptable to specific clinical tasks.

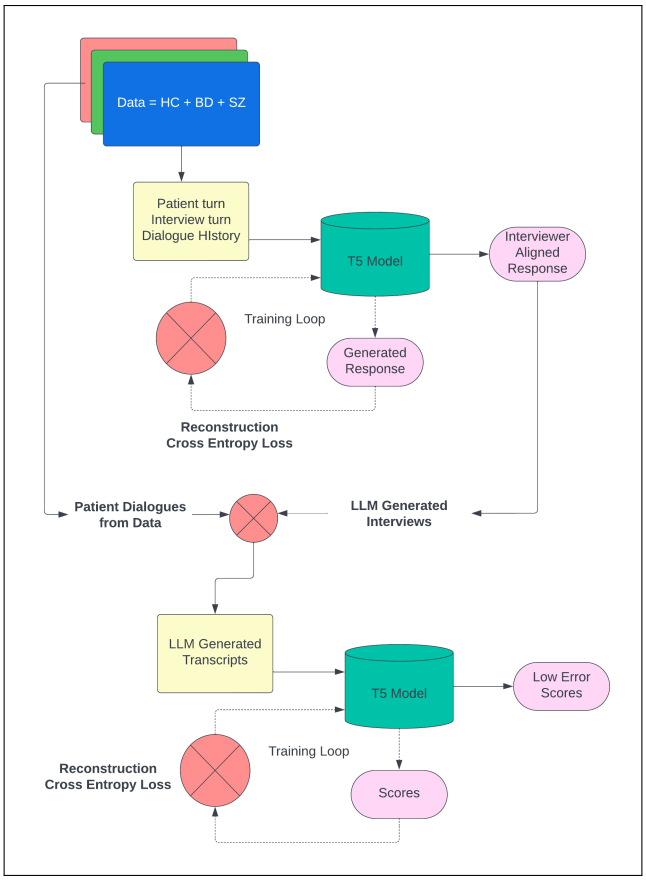

Figure 3: Chained Model Setup. Two standalone t5 models are chained by output and input. The Interview generator model works with patient dialogues to create LLM generated transcripts. This is fed into the score prediction model which outputs low error scores for the SSPA using a cross-entropy loss function.Picture resized for space limitations. Please zoom-in while reading review version.

BD Scene 1	BD Scene 2	SZ Scene 1	SZ Scene 2	HC Scene 1	HC Scene 2
0.0	0.06	0.03	0.04	0.2	0.04

Table 4: Difference of mean errors per case and scene.

Interest	Fluency	Clarity	Focus	Social
0.8	0.02	0.0	0.03	0.0

Table 5: Difference of mean errors per variable.

8 Limitations

In this study, we engaged 644 participants, which constitutes a relatively small sample size. While Patterson et al. (2001) originally identified eight variables in the SSPA, we selected only five for our analysis. The excluded variables were either unrelated to speech (e.g., personal grooming) or lacked expert raters due to their independence from the healthcare context (e.g., negotiation ability). We employed the T5 model for this task, primarily due to hardware constraints. Despite its smaller size, the T5 model demonstrated the capability to achieve relatively low error rates even with limited computational resources. This observation suggests that utilizing a larger model with more computational capacity could potentially reduce errors further. Furthermore, our study is limited to analyzing transcripts from audio recordings derived from the original dataset and does not incorporate multimodal aspects such as features of voice or audio.

Another limitation of our study concerns the use of a commercial language model, such as GPT-4/4o, exclusively for comparing annotations rather than conducting interviews. Although it could be argued that a commercial language model might

Scene/Class	GPT-4 Error	GPT-4o Error	Our Error	p 4	p - 4o
HC - Sc - 1	1.60	1.57	0.2	0.03	0.03
HC - Sc -2	1.70	1.66	0.04	0.03	0.03
SZ - Sc -1	1.44	1.50	0.03	0.03	0.03
SZ - Sc - 2	1.64	1.53	0.04	0.03	0.03
BD - Sc - 1	1.51	1.49	0.00	0.03	0.03
BD - Sc - 2	1.45	1.53	0.05	0.03	0.03

Table 6: Baseline Comparison with GPT4 and GPT4o. We can see that the p values comparing our error with GPT errors show a significant difference.

also be employed for interviews to compare performance outcomes, this approach raises significant ethical concerns. Firstly, most commercial language models do not possess adequate safeguards or specialized training to generate content that is safe for individuals with severe psychiatric conditions. Secondly, using such models could involve transmitting sensitive subject data and speech patterns to third-party systems, thereby raising serious ethical issues related to privacy and confidentiality. Consequently, commercial language models were restricted solely to annotation tasks using de-identified data.

The broader implications of these limitations merit careful consideration. Future work could explore how the currently excluded SSPA variables might be more formally defined and integrated into automated annotation pipelines using large language models (LLMs). While our study included 644 participants—a meaningful number in the context of psychiatric research—it remains relatively modest from the perspective of generalizability in AI applications for mental health. Nonetheless, this dataset represents one of the largest and medically validated corpora available for schizophrenia (SZ) and bipolar disorder (BD), laying essential groundwork for future model development. Due to hardware constraints, we employed the T5-base model, and it remains an open question whether scaling to larger variants (e.g., T5-XL or T5-XXL) would yield statistically significant performance improvements. We also limited the use of commercial models like GPT to annotation tasks only, in order to avoid exposing participants to unsupervised, third-party systems—particularly given the ethical concerns around deploying such models with vulnerable populations. Despite these constraints, our findings demonstrate that the speech patterns of trained psychiatrists can be reliably replicated with low error, opening the door to potential extensions such as modeling patient speech or augmenting data for low-resource clinical contexts. Finally, while we utilize established metrics such as BERTScore, ROUGE, cosine similarity, and RMSE, we acknowledge that these summarization benchmarks offer limited explainability with respect to individual-level communication quality. We encourage future work to incorporate more interpretable evaluation frameworks to deepen insight into both linguistic nuance and clinical relevance.

9 Ethical Concerns

This paper aims to demonstrate how modern language models can be deployed in clinical settings to collect and label data responsibly. We exclusively use labels that are well-established in clinical contexts. Importantly, this paper does not advocate for or implement the use of language models as diagnostic tools for mental health. We illustrate that markers of speech relevant to psychiatric healthcare can be predicted using language models. However, predicting variables like Interest or Focus should not be used or interpreted for unrelated tasks, such as advertising, targeted marketing, or any clinical purposes without appropriate expertise.

All data-related activities, including labeling, annotation, and sharing, were conducted with the approval of four independent academic Institutional Review Boards (IRBs). Participants in the original study provided informed consent. We adhere to all ethical codes established by the ACM and ACL. This paper involves numerous clinical experts in the labeling, adjudication, and language modeling processes, ensuring proper guidance and assistance. Using these models or concepts from this paper for non-clinical purposes or without expert guidance

in clinical contexts is strictly prohibited.

References

Ankit Aich, Avery Quynh, Varsha Badal, Amy Pinkham, Philip Harvey, Colin Depp, and Natalie Parde. 2022. Towards intelligent clinically-informed language analyses of people with bipolar disorder and schizophrenia. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2871–2887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Birnbaum, Sindhu Kiranmai Ernala, A. Rizvi, Elizabeth Arenare, Anna Van Meter, M. Choudhury, and J. Kane. 2019. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from facebook. *npj Schizophrenia*, 5.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

David Chandran, Deborah Robbins, Chin-Kuo Chang, Hitesh Shetty, Jyoti Sanyal, Johnny Downs, Marcella Fok, Michael Ball, Richard Jackson, Robert Stewart, Hannah Cohen, Jentien Vermeulen, Frederike Schirmbeck, Lieuwe Haan, and Richard Hayes. 2019. Use of natural language processing to identify obsessive compulsive symptoms in patients with schizophrenia, schizoaffective disorder or bipolar disorder. *Scientific Reports*, 9:1–7.

Nicholas Coombs, Wyatt Meriwether, James Caringi, and Sophia Newcomer. 2021. Barriers to healthcare access among u.s. adults with mental health challenges: A population-based study. *SSM - Population Health*, 15:100847.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–16, New York, NY, USA. Association for Computing Machinery.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.

Keith Harrigian and Mark Dredze. 2022. Then and now: Quantifying the longitudinal validity of self-disclosed depression diagnoses. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 59–75, Seattle, USA. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Keumhee Kang, Chanhee Yoon, and Eun Yi Kim. 2016. Identifying depressive users in twitter using multimodal analysis. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, pages 231–238, Los Alamitos, CA, USA. IEEE Computer Society.

Gina Kuperberg. 2010. Language in schizophrenia part 1: An introduction. *Language and linguistics compass*, 4:576–589.

Ellen Lee, John Torous, Munmun Choudhury, Colin Depp, Sarah Graham, Ho-Cheol Kim, Martin Paulus, John Krystal, and Dilip Jeste. 2021. Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tingting Liu, Salvatore Giorgi, Kenna Yadeta, H Andrew Schwartz, Lyle H Ungar, and Brenda Curtis. 2022. Linguistic predictors from facebook postings of substance use disorder treatment retention versus discontinuation. *The American journal of drug and alcohol abuse*, 48(5):573–585.

Brent Mausbach, Raeanne Moore, Christopher Bowie, Veronica Cardenas, and Thomas Patterson. 2008. A review of instruments for measuring functional recovery in those diagnosed with psychosis. *Schizophrenia bulletin*, 35:307–18.

Anne Merrill, Nicole Karcher, David Cicero, Theresa Becker, Anna Docherty, and John Kerns. 2017. Evidence that communication impairment in schizophrenia is associated with generalized poor task performance. *Psychiatry Research*, 249.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Metaicl: Learning to learn in context. *Preprint*, arXiv:2110.15943.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

Mahdieh Montazeri, Mitra Montazeri, Kambiz Bahaadinbeigy, Mohadeseh Montazeri, and Ali Afraz. 2022. Application of machine learning methods in predicting schizophrenia and bipolar disorders: A systematic review. *Health Science Reports*, 6.

María Palacios-Ariza, Esteban Morales-Mendoza, Jossie Murcia, Rafael Arias, Germán Lara-Castellanos, Andrés Cely-Jiménez, Juan Rincón-Acuña, Marcos Araúzo-Bravo, and Jorge McDouall. 2023. Prediction of patient admission and readmission in adults from a colombian cohort with bipolar disorder using artificial intelligence. *Frontiers in psychiatry*, 14:1266548.

Thomas L Patterson, Sherry Moscona, Christine L McKibbin, Kevin Davidson, and Dilip V Jeste. 2001. Social skills performance assessment among older patients with schizophrenia. *Schizophrenia Research*, 48(2):351–360.

Delaram Sadeghi, Afshin Shoeibi, Navid Ghassemi, Parisa Moridian, Ali Khadem, Roohallah Alizadehsani, Mohammad Teshnehlab, Juan Gorriz, and Saeid Nahavandi. 2021. An overview on artificial intelligence techniques for diagnosis of schizophrenia based on magnetic resonance imaging modalities: Methods, challenges, and future works.

Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 2018. Attention-based lstm for psychological stress detection from spoken language using distant supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6204–6208.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Task Description and Purpose

In this appendix we briefly describe the Social Skills Performance Assessment. We will talk about the purpose of the task, the task itself, and what we can gain from this task.

Task Motivation The Social Skills Performance Assessment, *abbrev. SSPA* is a mental-health instrument which serves as an indicator of social skill. The motivations, some of which we discussed in the introduction, is that people with psychiatric illnesses are more likely to show less cohesion and more disorganization in their speeches, as opposed to healthy control subjects. The SSPA task standardizes the way speech is measured for subjects with, and without psychiatric illnesses by having the subjects take on two tasks with expert clinicians.

For both tasks, the participants speak with a trained clinician. Their video and audio are recorded. Then transcribed. The labels mentioned in this paper were the clinicians rating the participants performance on the tasks to the two tasks.

Task Description There are two tasks to the SSPA. The first task is the neutral or friendly task, and the second task is the confrontational task.

The Friendly Task consists of the participant simulating a conversation as if they moved to a new neighborhood. They are asked to introduce themselves to the new neighbor. We observe people without psychiatric illnesses to briefly talk about 2-3 topics and stay consistent. People with illnesses tend to sway between 13-15 topics and are unable to concisely present thoughts.

The Confrontational Task consists of the participant complaining to their landlord after a leaky pipe has not been fixed for months. We observe that healthy controls are able to quickly articulate and talk only about the problem at hand. We observe that BD and SZ often talk about multiple different things and then talk about the problem given to them.

Task Outcomes Annotating clinical variables is a different task than classification. While these variables are not classifiers of psychiatric illnesses.

They are important features. These variables give clinicians and scientists much needed quantification in the field of life-long psychiatric illnesses. Therefore, it is imperative to bring modern technology to the equation and slowly make care and data collection accessible and efficient.

Prompt Details

This section describes the prompt that was used for GPT-4/4o to annotate the posts as described in Section §6.

> System Prompt - You are going to act as a clinical annotator. You will see a set of conversations between a doctor and a participant. You will also be told of a task. You need to return a python compatible list of five scores from a range of 1-5. Below I describe what these scores represent. Remember that for these scores 1 is lowest and 5 is highest.
>
> Interest - This score on 1-5 will describe how interested this person was in the conversation. Look at the participant's engagement in the conversation and rate this score.
>
> Fluency - This score on 1-5 will describe how fluent a person was. A person with more filler words will score lower.
>
> Clarity - This score on 1-5 will describe how clearly the subject was able to communicate their thoughts. A higher score shows better communication skills.
>
> Focus - This score on 1-5 will describe how concentrated the subject was on the task. A person who deviates off topic will score lower.
>
> Social Appropriateness - This score on 1-5 will describe how socially appropriate to the task this participant's score was. A higher score is more socially appropriate.
>
> Return the results as a list [] with five numbers for each of the scores above.
>
> User Prompt - In this task the participant has to introduce themselves as a new neighbor in the neighborhood.
>
> Doctor - Hey there, how are you? Participant - Hey I just moved. ...

Continues for entire conversation. The system prompt remains the same, while for each task the user-prompt changes.