Crisis prediction among tele-mental health patients: A large language model and expert clinician comparison

Christine Lee, Matthew Mohebbi, Erin O'Callahaghan, Mirène Winsberg

Submitted to: JMIR Mental Health
on: March 06, 2024

Table of Contents

Crisis prediction among tele-mental health patients: A large language model and expert clinician comparison

Christine Lee[1]; Matthew Mohebbi[1]; Erin O'Callahaghan[1]; Mirène Winsberg[1]

[1]Brightside Health San Francisco US

Corresponding Author:
Mirène Winsberg
Brightside Health
5214F Diamond Heights Blvd #3422
San Francisco
US

Abstract

Background: Due to recent advances in artificial intelligence (AI), large language models (LLMs) have emerged as a powerful tool for a variety of language related tasks, including sentiment analysis, and summarization of patient provided text. However, there is limited research on these models in the area of crisis prediction.

Objective: This study aimed to determine the performance of OpenAI's GPT-4 in predicting the likelihood of a mental health crisis episode based on patient provided information at intake among users of a national telemental health platform.

Methods: De-identified patient provided data was pulled from specific intake questions of the Brightside telehealth platform for 260 patients that later indicated they were experiencing suicidal ideation with a plan. 200 patients treated during the same time period who did not in the course of treatment endorse suicidal ideation were also randomly selected. Six Brightside clinicians (three psychologists and three psychiatrists) were shown patients' self-reported chief complaint and self-reported suicide attempt history but were blinded to the future course of treatment and other reported symptoms including suicidal ideation. They were asked a simple yes/no question regarding their prediction of suicidal ideation with plan along with a confidence level. GPT-4 was prompted with similar information and asked to provide answers to the same questions, enabling us to directly compare their performance.

Results: Overall accuracy (correctly assigning SI with plan vs no SI in the 460 examples) across the six raters using chief complaint alone ranged from 55.2% to 67% with an average of 62.1%. The GPT-4 based model had 61.5% accuracy. The addition of information regarding previous suicide attempts raised the average accuracy of the clinicians to 67.1% and GPT-4 to 67.0%. While overall performance of the GPT-4 based model approaches that of the clinicians, specificity was significantly lower than average for GPT-4 compared to clinicians in both scenarios (with and without history of previous suicide attempts). Average specificity across clinicians was 83.9% on chief complaint alone compared to 70.5% for the GPT-4 based model. Average specificity with the addition of the history of previous suicide attempts across the clinicians was 85.7% compared with 50.1% for the GPT-4 based model.

Conclusions: GPT-4 with a simple prompt design produced results on some metrics that approached that of a trained clinician. Additional work must be done before such a model could be piloted in a clinical setting. The model should undergo safety checks for bias given evidence that LLMs can perpetuate the biases of the underlying data they are trained upon. We believe that LLMs hold promise to augment identification of higher risk patients at intake and potentially deliver more timely care to patients.

(JMIR Preprints 06/03/2024:58129)
DOI: https://doi.org/10.2196/preprints.58129

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

✓ **Only make the preprint title and abstract visible.**
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Christine Lee, Matthew H. Mohebbi, Erin O'Callahaghan, Mirène Winsberg
All affiliated with Brightside Health

Crisis prediction among tele-mental health patients: A large language model and expert clinician comparison

Abstract

Background: Due to recent advances in artificial intelligence (AI), large language models (LLMs) have emerged as a powerful tool for a variety of language related tasks, including sentiment analysis, and summarization of provider-patient interactions. However, there is limited research on these models in the area of crisis prediction.

Objective: This study aimed to evaluate the performance of LLMs, specifically OpenAI's GPT-4, in predicting current and future mental health crisis episodes using patient provided information at intake among users of a national telemental health platform.

Methods:
De-identified patient provided data was pulled from specific intake questions of the Brightside telehealth platform, including the chief complaint, for 140 patients who indicated suicidal ideation (SI), and another 120 patients who later indicated SI with a plan during the course of treatment. Similar data was pulled for 200 randomly selected patients treated during the same time period who never endorsed SI. Six senior Brightside clinicians (three psychologists and three psychiatrists) were shown patients' self-reported chief complaint and self-reported suicide attempt history but were blinded to the future course of treatment and other reported symptoms including SI. They were asked a simple yes/no question regarding their prediction of endorsement of SI with plan along with their confidence level about the prediction. GPT-4 was provided similar information and asked to answer the same questions, enabling us to directly compare the performance of AI and clinicians.

Results: Overall, clinicians' average precision (0.698) was higher than GPT-4 (0.596) in identifying SI with plan at intake (n=140) vs. no SI (n=200) when using chief complaint alone, while sensitivity was higher for GPT-4 (0.621) than clinicians' average (0.529). The addition of suicide attempt history increased clinicians' average sensitivity (0.590) and precision (0.765), while increasing GPT-4 sensitivity (0.590) but decreasing GPT-4 precision (0.544). Performance decreased comparatively when predicting future SI with plan (n=120) vs no SI (n=200) with chief complaint only for clinicians (average sensitivity=0.399; average precision=0.594) and GPT-4 (sensitivity=0.458; precision=0.482). The addition of suicide attempt history increased performance comparatively for clinicians (average sensitivity=0.457; average precision=0.687) and GPT-4 (sensitivity=0.742; precision=0.476).

Conclusions: GPT-4 with a simple prompt design produced results on some metrics that approached that of a trained clinician. Additional work must be done before such a model could be piloted in a clinical setting. The model should undergo safety checks for bias given evidence that LLMs can perpetuate the biases of the underlying data they are trained upon. We believe that LLMs hold promise to augment identification of higher risk patients at intake and potentially deliver more timely care to patients.

Keywords: mental health; telehealth; PHQ-9; suicidal ideation; AI; LLM; OpenAI; GPT-4

Introduction

Suicide is a serious public health concern. Suicide rates have risen at an alarming rate in the past twenty years, and in the United States (U.S.), suicide is the second leading cause of death in adults aged 18-45 [1]. In 2021, approximately 50,000 people in the U.S. died by suicide, which marks the highest national rate of suicide in decades [2]. As suicide rates increase, the behavioral healthcare workforce in the U.S. has not expanded enough to keep up with these mental health demands, limiting the timely access to care that is essential for suicide risk detection, and prevention [3].

Suicide risk is difficult to predict. Research has demonstrated that there are numerous individual, relationship, community, and societal risk factors associated with suicide, such as history of previous suicide attempt, psychiatric diagnosis, sense of hopelessness, social isolation, community violence, and access to lethal means of suicide [4–9]. More recently, suicide theories and research suggest ideation-to-action pathways to help explain suicide risk; whereby people who think about suicide are at a higher risk to participate in suicidal behavior [10–13].

The prevalence of suicidal ideation (SI), which is defined as “thinking about, considering, or planning suicide” [14] is common, with 12.3 million Americans over the age 18 having thoughts of suicide in 2021[15]. SI is predictive of suicide attempts and completed suicide [16,17]. SI is also a more sensitive predictor of lifetime risk for suicide than imminent risk [18]. Research has suggested that among those exhibiting SI, there is a 29% conditional probability of making a suicide attempt [19]. Other research has shown that those with nearly daily SI were 5 to 8 times more likely to attempt suicide and 3 to 11 times more likely to die by suicide within 30 days [20].

Artificial Intelligence (AI) methods have been used for assessing mental health factors such as psychiatric symptom severity, diagnosis, and clinical risk using free text generated by the patient. Researchers utilizing natural language processing (NLP) and machine learning (ML) were able to identify suicidal behavior from electronic medical records [21], and detect suicidal ideation in a variety of different free-text settings [22]. In addition, an NLP-based system to determine likelihood of crisis in patient chat messages to their clinicians was developed and implemented with reliable retrospective and prospective performance as a clinical support tool for a crisis specialist team [23].

Recent advances in AI methods such as large language models (LLMs) have also shown success in a variety of medical applications. Both generalist LLMs such as GPT-4 and medical domain specific LLMs such as Med-PaLM 2 have exhibited medical competency on benchmarks such as the United States Medical Licensing Examination (USMLE) exam [24,25]. Generalist LLMs can sometimes outperform the domain specific LLMs, as was recently found with GPT-4 outperforming MedPaLM 2 on the MedQA medical benchmark [25]. Finally, Med-PaLM 2 was also found to be effective at determining psychiatric functioning from free text including patient generated information during patient interviews [26].

Objective

We seek to leverage the capabilities of LLMs to detect or predict suicidal ideation with plan among patients enrolled in a national telemental health platform, using patient-generated free text at intake.

We will benchmark the performance of this LLM-based prediction against a cohort of senior mental health clinician experts.

Methods

This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of WCG (Protocol #20240207).

Overview

The study consisted of clinicians completing a digital questionnaire where they were asked to predict whether a patient would endorse suicidal ideation (SI) with a plan during the course of their treatment, based on patient generated text describing their chief complaint. The same chief complaint texts were then served to a Large Language Model (LLM) GPT-4 with the same questionnaire instructions. The classification performance of the clinicians and GPT-4 were evaluated and compared.

Data Acquisition

The retrospective patient data used in this study were collected as part of standard of care at Brightside Health and de-identified for research purposes. All patients treated at Brightside consent at intake to terms of use and privacy policy which includes consenting to Brightside's use of their data for research purposes.

Inclusion Criteria

Data from patients who completed intake on the Brightside platform after March 15, 2023 and endorsed current SI (at intake) or subsequent SI (post-intake and during the course of treatment) were included in the study set, along with a random cohort of patients treated during the same time frame who never endorsed SI with plan. In order to be included in the study sample, patients had to attend at least one psychiatric or therapy appointment, and complete the chief complaint section of their digital intake form. Patients who left the chief complaint section empty were excluded.

Data and Outcome Variables

Patient generated free text (chief complaint) was extracted from patient intake as the answer to the question "In your own words, what are you feeling or experiencing?" and any personal identifiers (such as age, birthdate, names, location, email address, phone number and social security number) within the freetext was replaced with asterisks. Additional patient data extracted from intake included age, gender identity, and history of prior suicide attempt. Clinicians and the LLM did not have access to age or gender identity of the patients, and were only shown de-identified patient generated free text and then patient self-reported history of suicide attempts.

Suicidal ideation (SI) with plan was determined from answers to Question 9 of the Patient Health Questionnaire-9 (PHQ-9). The PHQ-9 is a self-report questionnaire consisting of nine questions measuring depression symptom severity ranging from 0 to 3 (not at all, several days, more than half the days, and nearly every day, respectively) within the past two weeks and includes a specific question related to the frequency of suicidal thoughts (item 9). If a patient endorses suicidal ideation on the Brightside platform (item 9 answer value > 0), a follow-up Brightside proprietary question

asks whether the suicidal thoughts are something the patient has made specific plans for. At Brightside, the PHQ-9 is administered to all patients at intake and requested every 2 weeks during the course of treatment. PHQ-9 answers at intake and at the date of first SI with plan relative to intake were also extracted for this study.

Classification Label Definitions

Patients positive for SI with plan were defined as those having endorsed suicidal ideation in the PHQ-9 at intake or at any point during the later course of treatment, and subsequently responded that the SI was something they had made specific plans for. Patients negative for SI with plan were defined as those with no PHQ-9 item 9 values > 0, i.e. those who had never endorsed SI in their PHQ-9 screenings.

Clinician Questionnaire Design

After creation of the study dataset, six clinicians employed at Brightside Health were recruited and all consented to participate in the study. Clinicians 1, 2, and 3 are senior psychiatrists with 18 to 30 years of clinical experience. Clinicians 4, 5, and 6 are senior psychologists with 10 to 23 years of clinical experience.

Each clinician was presented with the same questionnaire with the same randomized order of questions. Presented with a patient's de-identified chief complaint, they were asked to answer the following questions:

1. Do you expect this patient to endorse SI with a plan in the course of their treatment? (Yes/No)
2. How confident are you? (High/Medium/Low)

The clinicians were then presented with additional information as to whether the patient endorsed a previous suicide attempt and re-prompted with the same questions. All chief complaints and questions were displayed one at a time. Clinicians were given two weeks to complete the questionnaire with no other instructions or training, and were not required to complete it in a single sitting.

Large Language Model Prompt Design

In this study, we evaluated OpenAI's GPT-4 text only model's ability to predict whether a patient would endorse suicidal ideation with plan during the course of treatment. All model results were produced in Python with OpenAI's API and GPT-4 ("gpt-4-1106-preview" with default parameters) [27].

First the model context was constructed using step-by-step instructions and few-shot prompting with three real examples of chief complaints and their respective correct, expected answers (Textbox 1). Note that the examples chosen for the few-shot prompting were not included in the study dataset and were redacted for publication.

Next, the output format of the model was specified as JSON (JavaScript Object Notation) for ease of analysis. In addition to the prediction of suicidal ideation with plan during the course of treatment,

the model was also asked to provide a confidence level (High, Medium, Low) to the prediction (similar to the clinicians questionnaire) and to provide reasoning from a list of explicitly provided risk indicators.

Finally, the de-identified patient generated chief complaint text was given to the model in the user prompt. Each chief complaint was provided independently and then the LLM was reset back to the original context.

In order to evaluate the model's performance when served the additional information of patient self-reported prior suicide attempts, the sentence "I have attempted suicide before." or "I have never attempted suicided before." was appended to the end of the chief complaint and served as the prompt with the same context.

Textbox 1. GPT-4 few-shot prompt template for predicting suicidal ideation with plan in chief complaint text.

You are a trained psychiatric professional. You will be presented with a patient's answer to the question 'In your own words, what are you feeling or experiencing?', also referred to as a patient's chief complaint.

Your task is to determine based on a patient's chief complaint, if that patient will endorse suicidal ideation with plan in the course of their treatment.

You will generate your answer as JSON: {"at_risk": "", "confidence": "", "risk_indicators": "[]"}

Use the following step-by-step instructions to determine your answer:
Step 1. Answer in the key "at_risk" with only the values Yes or No
Step 2. Answer in the key "confidence" with only the values High, Medium, or Low, how confident you are in your answer
Step 3. Answer in the key "risk_indicators" with one or more values from the list of possible risk indicators below that determined your Yes or No answer.
Separate risk indicators with a "; ".
If Other, specify your answer as "Other-<insert your answer>". If no risk indicators are identified, answer with None.

List of possible risk indicators values:
Recent Stress, Loss, or Trauma
History of Trauma
Chronic medical conditions
Substance use
Previous suicide attempt
Lack or loss of relationships or support
Social isolation
Family history of suicide
Impulsive or aggressive language
Explicit mentions of suicide, suicidal thoughts, or self harm
Death imagery or metaphors
Apathy, indifference or emotional detachment

Sense of Hopelessness
Other

Here is an example of a chief complaint with a Yes to suicidal ideation with plan:
"<*text redacted for publication*> "
Your answer would be: {"risk_indicators": "Sense of Hopelessness; Social isolation; Explicit mentions of suicide, suicidal thoughts, or self harm", "at_risk": "Yes", "confidence": "High"}

Here is an example of a chief complaint with a No to suicidal ideation with plan: "<*text redacted for publication*>"
Your answer would be: {"risk_indicators": "None", "at_risk": "No", "confidence": "High"}

Here is an example of a chief complaint with a No to suicidal ideation with plan:
"<*text redacted for publication*>"
Your answer would be: {"risk_indicators": "None", "at_risk": "No", "confidence": "High"}

Performance Analysis

All analyses were performed in Python 3.8.12 with the package scikit-learn version 1.3.1 [28]. For comparison of performance, analyses were performed on positive for SI with plan at intake vs. negative for SI during the entire course of treatment as well as positive for SI with plan post-intake vs. the same data set of negative for SI during treatment.

Classification and Predictive Performance

Clinician and model performances in ability to predict whether a chief complaint text sample was positive for SI with plan, at intake and post intake, were evaluated for accuracy, sensitivity, specificity and precision. Accuracy was defined as the proportion of correctly predicted samples over the total number of samples. Precision (or positive predictive value) was defined as the proportion of correctly predicted positive samples over the total number of predicted positive samples. Sensitivity was defined as the proportion of correctly predicted positive samples over the total number of positive samples. Specificity was defined as the proportion of correctly predicted negative samples over the total number of negative samples. As an additional baseline reference, prior suicide attempt information (Yes or No) as a stand alone predictor was also included in the evaluation.

Clinician and LLM Agreement

To measure the agreement between clinician and GPT-4's predictions, Cohen's kappa statistic, which measures inter-rater agreement for categorical data, was calculated for each clinician and GPT-4 pairing.

Clinical Consensus and Confidence

Clinical consensus was defined as instances in which all clinicians answered with the same predicted outcome for a given sample, regardless of whether the prediction was correct. Rates of clinical consensus and rates of confidence were calculated to measure the variability and difficulty of clinical assessments on the given samples.

Accuracy of Clinical Consensus Influence on LLM Performance

To measure the influence of accuracy of clinical consensus on GPT-4 performance, subsets of chief complaint text samples where at least 1, 2, 3, 4, 5, or all 6 clinicians not only agreed but also correctly predicted the outcome for a given sample were evaluated for GPT-4 accuracy, sensitivity, specificity, and precision.

Risk Indicator Language and Clinician Performance

The GPT-4 prompt included a request to provide rationale for its prediction from a list of explicitly provided risk indicators (Textbox 1.). Clinician performance was then reevaluated on patient chief complaints with no GPT-4 identified risk indicators as a way to understand how difficult these cases were to clinical experts.

Due to the generative nature of a Large Language Model, GPT-4 occasionally will produce an answer that is not from the list of those that are explicitly defined in the instructions. For the purpose of this analysis, only the following explicit risk indicators as defined as exact string match were assessed: “Recent Stress, Loss, or Trauma”, “History of Trauma”, “Chronic medical conditions”, “Substance use”, “Previous suicide attempt”, “Lack or loss of relationships or support”, “Social isolation”, “Family history of suicide”, “Impulsive or aggressive language”, “Explicit mentions of suicide, suicidal thoughts, or self harm”, “Death imagery or metaphors”, “Apathy, indifference or emotional detachment”, and “Sense of Hopelessness”.

Results

At the date of study close (December 13, 2023) there were 260 patients who met inclusion criteria, and were positive for SI with plan. 140 patients were positive for SI with plan at time of intake and 120 patients were positive for SI with plan post-intake in their subsequent treatment. A random subset of 200 patients were selected from those who met inclusion criteria and were negative for SI with plan. Summary of the data can be found in Table 1.

Table 1. Summary of data for patients with no SI with plan (n=200), SI with plan indicated at intake (n=140), and SI with plan indicated post-intake (n=140). All results shown in average [95% Confidence Interval] or number of patients (%).

	No SI with Plan (n=200)	**SI with Plan At Intake (n=140)**	**SI with Plan Post Intake (n=120)**

Age	37.2 [35.7, 38.9]	34.4 [32.5, 36.3]	32.4 [30.3, 34.5]
Gender Identity			
Female	135 (67.5%)	76 (54.3%)	59 (49.2%)
Male	64 (32%)	57 (40.7%)	59 (49.2%)
Ethnicity			
White	152 (76.0%)	94 (67.1%)	73 (60.8%)
Hispanic	16 (8.0%)	20 (14.3%)	14 (11.7%)
Black	13 (6.5%)	13 (9.3%)	16 (13.3%)
Asian	10 (5.0%)	6 (4.3%)	8 (6.7%)
Other	9 (4.5%)	7 (5.0%)	9 (7.5%)
Average Chief Complaint Word Count	49.6 [41.3, 57.9]	58.0 [33., 83.1]	57.2 [44.2, 70.3]
Average Days between First SI with Plan Date and Chief Complaint	N/A	0.0 [nan, nan]	62.6 [52.4, 72.8]
Average PHQ-9 Total Score at First SI with Plan	N/A	21.1 [20.2, 21.9]	19.0 [17.8, 20.2]
# Patients with PHQ-9 Item 9 Score Value			
0			
1		32 (22.9%)	34 (28.3%)
2		34 (24.3%)	29 (24.2%)
3		74 (52.9%)	57 (47.5%)
With Specific Plan		140 (100.0%)	120 (100.0%)
Average PHQ-9 Total Score at Intake	13.5 [12.7, 14.2]	20.9 [20.1, 21.7]	18.3 [17.2, 19.4]
# Patients with PHQ-9 Item 9 Score Value			
0	200 (100%)	same as above	34 (28.3%)
1	0 (0%)		34 (28.3%)
2	0 (0%)		20 (16.7%)
3	0 (0%)		32 (26.7%)
With Specific Plan	0 (0%)		0 (0%)
Prior Suicide Attempt	14 (7%)	55 (39.3%)	40 (33.3%)

Prediction Performance

Predicting SI with Plan at Intake

The performance of prior suicide attempt alone to predict suicidal ideation with plan at the time of intake was similar to both GPT-4 and clinicians except for low sensitivity at 0.393 (Table 2).

GPT-4 performed with similar accuracy (0.671) and higher sensitivity (0.621) in predicting suicidal ideation with plan at the time of intake based on chief complaint text only, as compared to the

average accuracy (0.702) and sensitivity (0.529) across our 6 clinician participants (Table 2). However, GPT-4 performed with lower specificity (0.705) and precision (0.596) than average clinician specificity (0.823) and precision (0.698). The inter-rater agreement between GPT-4 and each clinician was moderate as indicated by an average Cohen's kappa of 0.488.

Additional knowledge of prior suicide attempt increased overall performance across clinicians (accuracy = 0.747; sensitivity = 0.590; specificity = 0.857; precision = 0.765). Additional knowledge of prior suicide attempt significantly increased sensitivity for GPT-4, but decreased accuracy, specificity and precision (accuracy = 0.644; sensitivity = 0.836; specificity = 0.51; precision = 0.544). The inter-rater agreement between GPT-4 and each clinician also decreased to an average Cohen's kappa of 0.386 with the additional information of prior suicide attempt.

Predicting SI with Plan Post-Intake

Performance decreased for both clinicians and GPT-4 when predicting future suicidal ideation with plan post-intake. Note that specificity results were consistent with predicting SI with plan at intake as there was no change in the negative samples..

GPT-4 performed with similar accuracy (0.612) and higher, but still poor, sensitivity (0.458) in predicting suicidal ideation with plan post-intake based solely on the chief complaintcompared to the average accuracy (0.664) and sensitivity (0.399) across the 6 clinicians (Table 2). GPT-4 performed with lower precision (0.482) than average clinician precision (0.594). The inter-rater agreement between GPT-4 and each clinician remained moderate at an average Cohen's kappa of 0.402.

Additional knowledge of prior suicide attempt increased performance across all clinicians (accuracy = 0.707; sensitivity = 0.457; precision = 0.687). Additional knowledge of prior suicide attempt significantly increased sensitivity for GPT-4, but decreased accuracy and precision (accuracy = 0.597; sensitivity = 0.742; precision = 0.476). The inter-rater agreement between GPT-4 and each clinician was lower with an average Cohen's kappa of 0.327 with the additional information.

Table 2. Performance results for predicting suicidal ideation with plan at time of intake and predicting suicidal ideation with plan in the future post-intake based on solely on chief complaintvs. chief complaint plus knowledge of prior attempt for GPT-4 and 6 clinicians. The performance of prior suicide attempt alone as a predictor is included for baseline reference.

SI with Plan At Intake (n=140) vs No SI with Plan (n=200)										
		True Negative	**False Positive**	**False Negative**	**True Positive**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	
	Baseline For Comparison: Prior Suicide Attempt Only	186	14	85	55	0.709	0.393	0.93	0.797	

		True Negative	False Positive	False Negative	True Positive	Accuracy	Sensitivity	Specificity	Precision	
CHIEF COMPLAINT TEXT ONLY	**GPT-4**	**141**	**59**	**53**	**87**	**0.671**	**0.621**	**0.705**	**0.596**	**Cohen's kappa with GPT-4**
	Clinician 1	160	40	58	82	0.712	0.586	0.8	0.672	0.534
	Clinician 2	189	11	95	45	0.688	0.321	0.945	0.804	0.363
	Clinician 3	138	62	48	92	0.676	0.657	0.69	0.597	0.559
	Clinician 4	183	17	85	55	0.7	0.393	0.915	0.764	0.436
	Clinician 5	162	38	58	82	0.718	0.586	0.81	0.683	0.497
	Clinician 6	156	44	52	88	0.718	0.629	0.78	0.667	0.538
					Average Across Clinicians	**0.702**	**0.529**	**0.823**	**0.698**	**0.488**
		True Negative	**False Positive**	**False Negative**	**True Positive**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	
CHIEF COMPLAINT TEXT + PRIOR SUICIDE ATTEMPT KNOWLEDGE	**GPT-4**	**102**	**98**	**23**	**117**	**0.644**	**0.836**	**0.51**	**0.544**	**Cohen's kappa with GPT-4**
	Clinician 1	163	37	49	91	0.747	0.65	0.815	0.711	0.464
	Clinician 2	194	6	89	51	0.721	0.364	0.97	0.895	0.21
	Clinician 3	152	48	39	101	0.744	0.721	0.76	0.678	0.499
	Clinician 4	187	13	67	73	0.765	0.521	0.935	0.849	0.329
	Clinician 5	173	27	53	87	0.765	0.621	0.865	0.763	0.399
	Clinician 6	159	41	47	93	0.741	0.664	0.795	0.694	0.415
					Average Across Clinicians	**0.747**	**0.590**	**0.857**	**0.765**	**0.386**
SI with Plan Post-Intake (n=120) vs No SI with Plan (n=200)										
		True Negative	**False Positive**	**False Negative**	**True Positive**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	
	Baseline For Comparison: Prior Suicide Attempt	*same results as at intake*		80	40	0.706	0.333	*same results as at intake*	0.741	

	Only									
		True Negative	**False Positive**	**False Negative**	**True Positive**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	
CHIEF COMPLAINT TEXT ONLY	**GPT-4**	*same results as at intake*		**65**	**55**	**0.612**	**0.458**	*same results as at intake*	**0.482**	**Cohen's kappa with GPT-4**
	Clinician 1			69	51	0.659	0.425		0.56	0.436
	Clinician 2			100	20	0.653	0.167		0.645	0.26
	Clinician 3			54	66	0.638	0.55		0.516	0.443
	Clinician 4			84	36	0.684	0.3		0.679	0.342
	Clinician 5			70	50	0.662	0.417		0.568	0.426
	Clinician 6			56	64	0.688	0.533		0.593	0.504
					Average Across Clinicians	**0.664**	**0.399**		**0.594**	**0.402**
		True Negative	**False Positive**	**False Negative**	**True Positive**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	
CHIEF COMPLAINT TEXT + PRIOR SUICIDE ATTEMPT KNOWLEDGE	**GPT-4**	*same results as at intake*		**31**	**89**	**0.597**	**0.742**	*same results as at intake*	**0.476**	**Cohen's kappa with GPT-4**
	Clinician 1			59	61	0.7	0.508		0.622	0.372
	Clinician 2			90	30	0.7	0.25		0.833	0.165
	Clinician 3			49	71	0.697	0.592		0.597	0.449
	Clinician 4			76	44	0.722	0.367		0.772	0.267
	Clinician 5			63	57	0.719	0.475		0.679	0.358
	Clinician 6			54	66	0.703	0.55		0.617	0.349
					Average Across Clinicians	**0.707**	**0.457**		**0.687**	**0.327**

Clinical Consensus and Confidence

Clinical consensus was defined as instances in which all 6 clinicians agreed on the predicted outcome for a given sample, regardless of whether the prediction was correct. Clinical consensus occurred in 52% of "no SI with plan" samples, 40.7% of "SI with plan at intake" samples, and 40% of "SI with plan post-intake" samples (Table 3). For SI with plan samples with a clinical consensus, the agreed upon prediction was correct 61.4% of the time for "SI with plan at intake" vs much lower at 25% of the time for "SI with plan post-intake". For "no SI with plan" samples, the clinician agreed upon prediction was correct at a high rate of 98.1%.

Additionally, clinicians, on average, had lower rates of high confidence (even when answers were

correct) compared to GPT-4 (Table 4). On average, clinicians answered correctly “No with High Confidence” in 9.5% of “no SI with plan” samples vs. GPT-4 answered “No with High Confidence” in 35%. Clinicians answered correctly “Yes with High Confidence” in 15.7% of “SI with plan at intake” samples vs. GPT-4 at 29.3%. Rates of correctly answered “Yes with High Confidence” were lower in “SI with plan post-intake” samples, but were higher for GPT-4 compared to average clinician rates (13.3% vs 7.2%).

Table 3. Rates of clinical consensus, defined as instances in which all 6 clinicians agreed on the predicted outcome for a given sample

	No SI with Plan (n=200)	SI with Plan At Intake (n=140)	SI with Plan Post-Intake (n=120)
# Samples with Clinical Consensus	104 (52%)	57 (40.7%)	48 (40%)
Clinical Consensus Predicted SI with Plan	2 (1.9%)	35 (61.4%)	12 (25%)
Clinical Consensus Predicted No SI with Plan	102 (98.1%)	22 (38.6%)	36 (75%)

Table 4. Rates of High Confidence Answers

	No SI with Plan (n=200)		SI w Plan At Intake (n=140)		SI w Plan Post-Intake (n=120)	
	Answered Yes with High Confidence	**Answered No with High Confidence**	**Answered Yes with High Confidence**	**Answered No with High Confidence**	**Answered Yes with High Confidence**	**Answered No with High Confidence**
Clinician 1	5 (2.5%)	6 (3%)	45 (32.1%)	1 (0.7%)	16 (13.3%)	2 (1.7%)
Clinician 2	0 (0%)	19 (9.5%)	5 (3.6%)	7 (5.0%)	1 (0.8%)	4 (3.3%)
Clinician 3	2 (1%)	41 (20.5%)	20 (14.3%)	9 (6.4%)	9 (7.5%)	6 (5.0%)
Clinician 4	0 (0%)	0 (0%)	1 (0.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Clinician 5	0 (0%)	2 (1%)	23 (16.4%)	0 (0.0%)	5 (4.2%)	3 (2.5%)
Clinician 6	2 (1%)	46 (23%)	38 (27.1%)	13 (9.3%)	21 (17.5%)	12 (10.0%)
Average Across Clinicians	**1.5 (0.75%)**	**19 (9.5%)**	**22 (15.7%)**	**5 (3.6%)**	**8.7 (7.2%)**	**4.5 (3.8%)**

GPT-4	1 (0.5%)	70 (35.0%)	41 (29.3%)	17 (12.1%)	16 (13.3%)	14 (11.7%)

Accuracy of Clinical Consensus and GPT-4 Performance

A range of accurate clinical consensus samples was defined as samples where a number of clinicians, ranging from at least 1 to all 6, not only agreed on the predicted outcome but also correctly predicted the outcome. There were 316 samples of the “SI with plan at intake” and “no SI with plan” samples where at least one clinician predicted the outcome correctly vs 137 samples where all 6 clinicians predicted the outcome correctly (Table 5). There were 282 samples of the “SI with plan post-intake” and “no SI with plan” samples where at least one clinician predicted the outcome correctly vs 114 samples where all 6 clinicians predicted the outcome correctly.

As the accurate clinical consensus threshold increased, GPT-4 performance increased significantly on those samples (Table 5). When assessing “SI with plan at intake” and “no SI with plan” samples with a clinical consensus of 3 or more and correct predictions, GPT-4 performed with an accuracy of 0.811, sensitivity of 0.912, specificity of 0.765, and precision of 0.635. When assessing “SI with plan post-intake” and “no SI with plan” samples with a clinical consensus of 3 or more and correct predictions, GPT-4 performed with an accuracy of 0.801, sensitivity of 0.857, and precision of 0.512.

Table 5. Performance results for GPT-4 solely on chief complaint in samples where at least 1, 2, 3, 4, 5, or all 6 clinicians correctly predicted the outcome of those samples.

SI with Plan At Intake (original n=140) vs No SI with Plan (original n=200)									
# Clinicians Correctly Predicted Samples Consensus Threshold	**# Samples**	**True Negative**	**False Positive**	**False Negative**	**True Positive**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
>=1	316	141	57	32	86	0.718	0.729	0.712	0.601
>=2	284	141	52	14	77	0.768	0.846	0.731	0.597
>=3	259	137	42	7	73	0.811	0.912	0.765	0.635
>=4	236	133	36	2	65	0.839	0.97	0.787	0.644
>=5	200	123	24	0	53	0.88	1	0.837	0.688
6	137	89	13	0	35	0.905	1	0.873	0.729
SI with Plan Post-Intake (original n=120) vs No Si with Plan (original n=200)									
# Clinicians Correctly Predicted Samples	**# Samples**	**True Negative**	**False Positive**	**False Negative**	**True Positive**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**

Consensus Threshold								
>=1	282	*same as above*		31	53	0.688	0.631	*same as above*	0.482
>=2	266			23	50	0.718	0.685		0.49
>=3	233			10	44	0.777	0.815		0.512
>=4	211			6	36	0.801	0.857		0.5
>=5	169			1	21	0.852	0.955		0.467
6	114			0	12	0.886	1		0.48

Risk Indicators Identified in Chief Complaint Text by GPT-4

At least one risk indicator was identified in the chief complaint text by GPT-4 on 45.5% of “No SI with plan” samples (Table 6). 70% of “SI with plan at intake” samples and 54.2% of “SI with plan post-intake” samples had at least one GPT_4 identified risk indicator. The most common risk indicator in “SI with plan at intake” samples identified by GPT-4 was “sense of hopelessness” (in 40.0% of samples, compared to 27.5% of “SI with plan post-intake” and 16.5% of “No SI with plan”. The most common risk indicator in “No SI with plan” samples was “recent stress, loss or trauma” (in 25.5% of samples, compared to 22.1% of “SI with plan at intake” samples and 17.5% of “SI with plan post-intake” samples. In addition, the rate of identification of “social isolation” as a risk factor in “SI with plan post-intake” samples (12.5%) was higher than in both “No SI with plan” (5.7%) samples and “SI with plan at intake” samples (6.5%)

Table 6. Number of samples per explicit risk indicator identified by GPT-4

# of Risk Indicators Identified by GPT-4	**No SI with Plan (n=200)**	**SI with Plan At Intake (n=140)**	**SI with Plan Post-Intake (n=120)**
0	109 (54.5%)	42 (30.0%)	55 (45.8%)
1	34 (17.0%)	28 (20.0%)	22 (18.3%)
2	34 (17.0%)	37 (26.4%)	18 (15.0%)
3	16 (8.0%)	22 (15.7%)	15 (12.5%)
4	4 (2.0%)	6 (4.3%)	8 (6.7%)
5	3 (1.5%)	3 (2.1%)	1 (0.8%)
6	0 (0%)	2 (1.4%)	1 (0.8%)
Risk Indicator Identified by GPT-4	**No SI with Plan (n=200)**	**SI with Plan At Intake (n=140)**	**SI with Plan Post Intake (n=120)**
Sense of Hopelessness	33 (16.5%)	56 (40.0%)	33 (27.5%)
Explicit mentions of suicide, suicidal	2 (1%)	38 (27.1%)	19 (15.8%)

thoughts, or self harm			
Recent Stress, Loss, or Trauma	51 (25.5%)	31 (22.1%)	21 (17.5%)
Apathy, indifference or emotional detachment	19 (9.5%)	22 (15.7%)	19 (15.8%)
Lack or loss of relationships or support	22 (11%)	17 (12.1%)	12 (10.0%)
Social isolation	13 (6.5%)	8 (5.7%)	15 (12.5%)
Chronic medical conditions	13 (6.5%)	13 (9.3%)	8 (6.7%)
History of Trauma	13 (6.5%)	10 (7.1%)	8 (6.7%)
Impulsive or aggressive language	3 (1.5%)	8 (5.7%)	6 (5.0%)
Previous suicide attempt	0 (0%)	9 (6.4%)	1 (0.8%)
Substance use	10 (5%)	6 (4.3%)	3 (2.5%)
Family history of suicide	0 (0%)	0 (0.0%)	1 (0.8%)
Death imagery or metaphors	2 (1%)	1 (0.7%)	0 (0.0%)

Chief Complaints with No Risk Indicators and Clinician Performance

Assessing clinicians' performance on samples where GPT-4 identified no explicit risk indicators in the chief complaint text, the average clinician sensitivity was found to be low for both "SI with plan at intake" and "SI with plan post-intake" –0.222 and 0.170, respectively (Table 7). Average clinician specificity and precision were high for both "SI with plan at intake" and "SI with plan post-intake" at 0.925 and 0.626 vs. 0.925 and 0.598, respectively. While the sample size in this analysis was significantly decreased n=109 [54.5%] for "No SI with plan", n=42 [30%] for "SI with plan at intake" and n=55 [45.8%] for "SI with plan post-intake", clinicians' performance resulted in fewer false positives and a lower rate of positive prediction, indicating that clinicians are less likely to predict suicidal ideation with plan in patients where GPT did not identify any risk factors.

Table 7. Performance results for chief complaint text only samples where GPT-4 identified zero explicit risk indicators.

SI with Plan At Intake (n=42) vs No SI with Plan (n=109)

	True Negative	False Positive	False Negative	True Positive	Accuracy	Sensitivity	Specificity	Precision
GPT-4	**109**	**0**	**40**	**2**	**0.735**	**0.048**	**1**	**1**
Clinician 1	100	9	33	9	0.722	0.214	0.917	0.5
Clinician 2	108	1	40	2	0.728	0.048	0.991	0.667
Clinician 3	91	18	27	15	0.702	0.357	0.835	0.455
Clinician 4	109	0	39	3	0.742	0.071	1	1
Clinician 5	101	8	28	14	0.762	0.333	0.927	0.636
Clinician 6	96	13	29	13	0.722	0.31	0.881	0.5
				Average Across Clinicians	**0.730**	**0.222**	**0.925**	**0.626**
SI with Plan Post-Intake (n=55) vs No SI with Plan (n=109)								
	True Negative	**False Positive**	**False Negative**	**True Positive**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
GPT-4	**109**	**0**	**54**	**1**	**0.671**	**0.018**		**1**
Clinician 1	100	9	45	10	0.671	0.182		0.526
Clinician 2	108	1	54	1	0.665	0.018		0.5
Clinician 3	91	18	37	18	0.665	0.327	*same as above*	0.5
Clinician 4	109	0	51	4	0.689	0.073		1
Clinician 5	101	8	44	11	0.683	0.2		0.579
Clinician 6	96	13	43	12	0.659	0.218		0.48
				Average Across Clinicians	**0.672**	**0.170**		**0.598**

Discussion

The objective of this study was to evaluate the performance of the foundation large language model (LLM) GPT-4 compared to experienced mental health clinicians in predicting suicidal ideation with plan based on a patient-generated chief complaint free text at intake on a national telemental health platform. This study supports previous research that LLMs are able to perform comparably to

clinicians in medical applications, and that generalist models like GPT-4 are able to deliver comparable performance without specialized fine-tuning or domain expertise [24,25].

Results

GPT-4 is capable of predicting risk of suicidal ideation with plan using patient generated chief complaint free text without extensive work on prompt design and without being trained explicitly on this task. The performance of these GPT-4-based predictions approach those of the clinicians on a variety of measures.

The variability in clinicians performance and agreement indicates that identifying suicidal ideation with plan in patient text alone is a difficult problem even for clinical experts. However, using the clinical experts in this study as a benchmark, GPT-4 was still able to perform comparably in sensitivity but with lower specificity and precision. When assessing GPT-4 on samples with high clinician agreement and performance, this study found that GPT-4 was capable of significantly high sensitivity as well as specificity. These results support that models such as GPT-4, without large amounts of time spent on highly complex data cleaning or model training, are capable of identifying risk of crisis comparable to the average clinician.

This study also explored the utilization of GPT-4 as a natural language processing (NLP) technique for the extraction of meaningful clinical information. GPT-4 was able to identify and return explicit indicators of risk in text, such as sense of hopelessness, that could further assist in crisis triaging and resourcing.

In addition, while not a specific aim or analysis in this study, the average clinician took approximately 3 hours to evaluate the 460 samples of text provided. GPT-4 completed the full evaluation in less than 10 minutes, without optimization for compute or memory, highlighting the possible increased operational efficiency that could be leveraged by automating a tedious and emotionally trying manual task.

Taking into consideration the current behavioral healthcare workforce shortage, and the increasing rates of suicide, there is a need for scalable, efficient, technology-enabled screening techniques such as the one utilized in this study, to assist with suicide risk detection. More efficient risk detection will allow for faster delivery of interventions to help prevent suicide attempts. Utilization of technology for this purpose would also be a cost-saving, and efficient way to more broadly screen for suicide risk. Patients deemed at high risk might be triaged to clinicians with greater expertise in managing suicidality.

Responsible integration and use of generative as a screening tool for predicting the likelihood of crisis would depend on achieving at least similar accuracy to a team of clinicians, and should always follow-up with clinician review who would be given additional context behind the GPT-4 based prediction and have access to additional clinical data.

Overall, GPT-4 shows promise as a solution to help clinicians deliver more timely care.

Limitations

We do not intend for this study, the LLM choice, or the prompt design to be viewed as a

generalizable solution to predict and identify suicidal risk. Instead we have shown how the capabilities of these LLMs can be tailored to specific psychiatric assessments, and how they compare to the limitations of expert clinician predictions. We hope that the findings encourage further research.

A number of limitations in this study must be addressed before the results of such a system could be applied in practice, including but not limited to data from a larger and/or more diverse population, use of other LLMs and in-particular LLMs that were built for application in the medical domain and a greater exploration of prompt design and its impact on performance. Similar to the use of real-time clinical decision support for precision prescribing at Brightside that is reliant on medical decision making by trained clinicians, the use of LLM for triage would be limited to suggestions and distillation of information for further clinician assessment [29].

Suicide has been notoriously difficult to predict. Due to the difficult nature of identifying or predicting future SI with plan, precision uncertainties are a reality of treating higher severity behavioral health patients as can be seen by the number of false positives and lower precision across several clinicians. Due to this uncertainty, awareness of risk does not necessarily dictate treatment decisions, but might influence triage to a provider with more expertise in treating suicidality.

GPT-4 was on the higher end for false positives with chief complaint text only relative to the clinicians and when prior attempt knowledge was added this rate was almost doubled this metric relative to the worst performing clinician. While work should be done to further align this GPT-4 based system with the expert clinicians, especially with prior attempt information, these false positives are clearly a reality in treating patients today.

GPT-4 was on the lower end for false negatives relative to the clinicians, in some cases having half as many false positives as the worst performing clinicians. It is our view that increasing awareness around potential risk through use of systems such as this is valuable, especially for clinicians that have less expertise.

Finally, as previously discussed, LLMs have tendencies to perpetuate biases inherent in the data on which they are trained [30]. Future work should explore how these biases may influence the quality of the prediction within different subpopulations of patients [31].

Conclusions

The use of machine learning and LLMs to analyze speech and language patterns offers an opportunity for behavioral health clinicians and researchers to explore technologies such as these to assist with detection and prediction of mental health conditions, along with specific symptoms such as suicidal thoughts, intent, and behaviors [32]. This study served as a model for comparing the predictive value of generative AI to clinician (imperfect) predictions when both were given access to the same limited data set. Research evaluating applications of AI technology to human speech, language, and behavior is in its infancy, but findings such as the ones presented in this study may help clinicians and researchers leverage the potential of LLMs to help those struggling with mental illness. Generative AI has the potential to transform areas of mental health care that might otherwise be overlooked. However, great care must be taken by both developers of this technology and the clinicians who deploy them to ensure that the benefits far outweigh the safety challenges and risks.

Further research is encouraged in this area, with consideration of ethical and clinical implications of the use of AI for detecting and predicting mental health issues [32]. This research will assist in

setting standards and guidelines for how such use could be deployed.

Acknowledgements

The authors would like to thank the six clinicians who contributed their time to participate in data collection for this study.

Conflicts of Interest

The authors all hold stock in Brightside Health, Inc. and are all employees of Brightside Health, Inc. The authors declare that this study received funding from Brightside Health. The authors are employees of Brightside Health, but aside from employment status, the funder was not involved in the study design, interpretation of data, or the decision to submit for publication.

Abbreviations

AI: Artificial intelligence
LLM: Large language model
GPT: Generative pre-training transformer
ML: Machine learning
NLP: Natural language processing
SI: Suicidal ideation

References

1. WISQARS (Web-based Injury Statistics Query and Reporting System) | Injury Center | CDC. 2023. Available from: https://www.cdc.gov/injury/wisqars/index.html [accessed Jan 20, 2024]
2. Suicide Data and Statistics | Suicide Prevention | CDC. 2023. Available from: https://www.cdc.gov/suicide/suicide-data-statistics.html [accessed Jan 20, 2024]
3. Understanding the U.S. Behavioral Health Workforce Shortage. 2023. doi: 10.26099/5km6-8193
4. Beautrais AL, Joyce PR, Mulder RT, Fergusson DM, Deavoll BJ, Nightingale SK. Prevalence and comorbidity of mental disorders in persons making serious suicide attempts: a case-control study. Am J Psychiatry 1996 Aug;153(8):1009–1014. PMID:8678168
5. Risk and Protective Factors | Suicide Prevention | CDC. 2023. Available from: https://www.cdc.gov/suicide/factors/index.html [accessed Jan 20, 2024]
6. Ribeiro JD, Huang X, Fox KR, Franklin JC. Depression and hopelessness as risk factors for suicide ideation, attempts and death: meta-analysis of longitudinal studies. Br J Psychiatry 2018 May;212(5):279–286. doi: 10.1192/bjp.2018.27
7. Motillon-Toudic C, Walter M, Séguin M, Carrier J-D, Berrouiguet S, Lemey C. Social isolation and suicide risk: Literature review and perspectives. Eur Psychiatry 2022;65(1):e65. doi: 10.1192/j.eurpsy.2022.2320
8. Castellví P, Miranda-Mendizábal A, Parés-Badell O, Almenara J, Alonso I, Blasco MJ, Cebrià A, Gabilondo A, Gili M, Lagares C, Piqueras JA, Roca M, Rodríguez-Marín J, Rodríguez-Jimenez T, Soto-Sanz V, Alonso J. Exposure to violence, a risk for suicide in youths and young adults. A meta-analysis of longitudinal studies. Acta Psychiatr Scand 2017 Mar;135(3):195–

211. doi: 10.1111/acps.12679
9. Irigoyen M, Porras-Segovia A, Galván L, Puigdevall M, Giner L, De Leon S, Baca-García E. Predictors of re-attempt in a cohort of suicide attempters: A survival analysis. J Affect Disord 2019 Mar 15;247:20–28. doi: 10.1016/j.jad.2018.12.050
10. Hubers A a. M, Moaddine S, Peersmann SHM, Stijnen T, van Duijn E, van der Mast RC, Dekkers OM, Giltay EJ. Suicidal ideation and subsequent completed suicide in both psychiatric and non-psychiatric populations: a meta-analysis. Epidemiol Psychiatr Sci 2018 Apr;27(2):186–198. PMID:27989254
11. Jobes DA, Joiner TE. Reflections on Suicidal Ideation. Crisis 2019 Jul;40(4):227–230. PMID:31274031
12. Nock MK, Borges G, Bromet EJ, Cha CB, Kessler RC, Lee S. Suicide and Suicidal Behavior. Epidemiol Rev 2008;30(1):133–154. PMID:18653727
13. O'Connor RC, Kirtley OJ. The integrated motivational–volitional model of suicidal behaviour. Philos Trans R Soc B Biol Sci 2018 Sep 5;373(1754):20170268. PMID:30012735
14. Suicide - National Institute of Mental Health (NIMH). Available from: https://www.nimh.nih.gov/health/statistics/suicide [accessed Jan 20, 2024]
15. Suicide statistics. Am Found Suicide Prev. Available from: https://afsp.org/suicide-statistics/ [accessed Jan 20, 2024]
16. Simon GE, Rutter CM, Peterson D, Oliver M, Whiteside U, Operskalski B, Ludman EJ. Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? Psychiatr Serv Wash DC 2013 Dec 1;64(12):1195–1202. PMID:24036589
17. Simon GE, Yarborough BJ, Rossom RC, Lawrence JM, Lynch FL, Waitzfelder BE, Ahmedani BK, Shortreed SM. Self-Reported Suicidal Ideation as a Predictor of Suicidal Behavior Among Outpatients With Diagnoses of Psychotic Disorders. Psychiatr Serv American Psychiatric Publishing; 2019 Mar;70(3):176–183. doi: 10.1176/appi.ps.201800381
18. Ursano RJ, Heeringa SG, Stein MB, Jain S, Raman R, Sun X, Chiu WT, Colpe LJ, Fullerton CS, Gilman SE, Hwang I, Naifeh JA, Nock MK, Rosellini AJ, Sampson NA, Schoenbaum M, Zaslavsky AM, Kessler RC. Prevalence and correlates of suicidal behavior among new soldiers in the U.S. Army: results from the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Depress Anxiety 2015 Jan;32(1):3–12. PMID:25338964
19. Rossom RC, Coleman KJ, Ahmedani BK, Beck A, Johnson E, Oliver M, Simon GE. Suicidal ideation reported on the PHQ9 and risk of suicidal behavior across age groups. J Affect Disord 2017 Jun;215:77–84. PMID:28319695
20. Stone M, Laughren T, Jones ML, Levenson M, Holland PC, Hughes A, Hammad TA, Temple R, Rochester G. Risk of suicidality in clinical trials of antidepressants in adults: analysis of proprietary data submitted to US Food and Drug Administration. BMJ 2009 Aug 11;339:b2880. PMID:19671933
21. Carson NJ, Mullin B, Sanchez MJ, Lu F, Yang K, Menezes M, Cook BL. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PloS One 2019;14(2):e0211116. PMID:30779800
22. Arowosegbe A, Oyelade T. Application of Natural Language Processing (NLP) in Detecting and Preventing Suicide Ideation: A Systematic Review. Int J Environ Res Public Health 2023 Jan 13;20(2):1514. PMID:36674270
23. Swaminathan A, López I, Mar RAG, Heist T, McClintock T, Caoili K, Grace M, Rubashkin M, Boggs MN, Chen JH, Gevaert O, Mou D, Nock MK. Natural language processing system for rapid detection and intervention of mental health crisis chat messages. Npj Digit Med 2023 Nov 21;6(1):213. doi: 10.1038/s41746-023-00951-3
24. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A,

Mansfield P, Demner-Fushman D, Agüera Y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. Nature 2023 Aug 3;620(7972):172–180. doi: 10.1038/s41586-023-06291-2

25. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, King N, Larson J, Li Y, Liu W, Luo R, McKinney SM, Ness RO, Poon H, Qin T, Usuyama N, White C, Horvitz E. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. arXiv; 2023. Available from: http://arxiv.org/abs/2311.16452 [accessed Dec 4, 2023]
26. Galatzer-Levy IR, McDuff D, Natarajan V, Karthikesalingam A. The Capability of Large Language Models to Measure Psychiatric Functioning.
27. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A-L, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser Ł, Kamali A, Kanitscheider I, Keskar NS, Khan T, Kilpatrick L, Kim JW, Kim C, Kim Y, Kirchner H, Kiros J, Knight M, Kokotajlo D, Kondraciuk Ł, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li CM, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney SM, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O'Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, Peres F de AB, Petrov M, Pinto HP de O, Michael, Pokorny, Pokrass M, Pong V, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such FP, Summers N, Sutskever I, Tang J, Tezak N, Thompson M, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe JFC, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang JJ, Wang A, Wang B, Ward J, Wei J, Weinmann CJ, Welihinda A, Welinder P, Weng J, Weng L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S, Zheng T, Zhuang J, Zhuk W, Zoph B. GPT-4 Technical Report. arXiv; 2023. doi: 10.48550/arXiv.2303.08774
28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D. Scikit-learn: Machine Learning in Python. Mach Learn PYTHON.
29. O'Callaghan E, Sullivan S, Gupta C, Belanger HG, Winsberg M. Feasibility and acceptability of a novel telepsychiatry-delivered precision prescribing intervention for anxiety and depression. BMC Psychiatry 2022 Dec;22(1):483. doi: 10.1186/s12888-022-04113-9
30. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias,

ethics, limitations and future scope. Internet Things Cyber-Phys Syst 2023 Jan 1;3:121–154. doi: 10.1016/j.iotcps.2023.04.003

31. Timmons AC, Duong JB, Simo Fiallo N, Lee T, Vo HPQ, Ahle MW, Comer JS, Brewer LC, Frazier SL, Chaspari T. A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health. Perspect Psychol Sci J Assoc Psychol Sci 2023 Sep;18(5):1062–1096. PMID:36490369
32. Diaz-Asper C, Hauglid MK, Chandler C, Cohen AS, Foltz PW, Elvevåg B. A framework for language technologies in behavioral research and clinical applications: Ethical challenges, implications, and solutions. Am Psychol 2024 Jan;79(1):79–91. doi: 10.1037/amp0001195