Psychiatry Research 341 (2024) 116119

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Assessing dimensions of thought disorder with large language models: The tradeoff of accuracy and consistency

Samuel L. Pugh [a,b], Chelsea Chandler [b], Alex S. Cohen [c,d], Catherine Diaz-Asper [e], Brita Elvevåg [f,g,*], Peter W. Foltz [b]

[a] Department of Computer Science, University of Colorado Boulder, United States
[b] Institute of Cognitive Science, University of Colorado Boulder, United States
[c] Department of Psychology, Louisiana State University, United States
[d] Center for Computation and Technology, Louisiana State University, United States
[e] Department of Psychology, Marymount University, United States
[f] Department of Clinical Medicine, University of Tromsø–The Arctic University of Norway, Norway
[g] Norwegian Center for Clinical Artificial Intelligence, University Hospital of North Norway, Norway

ARTICLE INFO

Keywords:
Natural language processing
Schizophrenia
Speech
Language
Incoherence
LLM
Psychiatry

ABSTRACT

Natural Language Processing (NLP) methods have shown promise for the assessment of formal thought disorder, a hallmark feature of schizophrenia in which disturbances to the structure, organization, or coherence of thought can manifest as disordered or incoherent speech. We investigated the suitability of modern Large Language Models (LLMs - e.g., GPT-3.5, GPT-4, and Llama 3) to predict expert-generated ratings for three dimensions of thought disorder (coherence, content, and tangentiality) assigned to speech samples collected from both patients with a diagnosis of schizophrenia ($n = 26$) and healthy control participants ($n = 25$). In addition to (1) evaluating the accuracy of LLM-generated ratings relative to human experts, we also (2) investigated the degree to which the LLMs produced consistent ratings across multiple trials, and we (3) sought to understand the factors that impacted the consistency of LLM-generated output. We found that machine-generated ratings of the level of thought disorder in speech matched favorably those of expert humans, and we identified a tradeoff between accuracy and consistency in LLM ratings. Unlike traditional NLP methods, LLMs were not always consistent in their predictions, but these inconsistencies could be mitigated with careful parameter selection and ensemble methods. We discuss implications for NLP-based assessment of thought disorder and provide recommendations of best practices for integrating these methods in the field of psychiatry.

1. Introduction

Thought disorder is a major component in the overall presenting phenomenology in schizophrenia, and the presence and characterization of this thought disorder is important in differential diagnosis. Its presence is inferred from the speech of patients, which may be of reduced intelligibility and be increasingly disorganized such that it is difficult, and sometimes impossible, for the listener to comprehend. This genre of speech is then inferred to be indicative of disturbances in the structure, organization, and coherence of the assumed underlying thought processes. Speech, therefore, provides an important modality for assessing thought disorder (Andreasen and Grove, 1986; Corcoran and Cecchi, 2020; DeLisi, 2001). Thought disorder is generally evaluated via observation of speech during a clinical examination, and formally measured using semi-structured scales such as the Thought and Language Disorder (TALD) scale (Kircher et al., 2014) or the Scale for Assessment of Thought, Language, and Communication (TLC) (Andreasen, 1986). These scales provide a structured procedure to evaluate speech for signs of thought disorder, however they are time-consuming to administer, require extensive training to conduct, and can suffer from suboptimal inter-rater reliability in practice.

In contrast to human-administered assessments, computerized natural language processing (NLP) tools provide the possibility of automatically and consistently quantifying elements of speech (e.g., incoherence) indicative of thought disorder if the speech prompts and study design are optimized (Elvevåg et al., 2017; Foltz et al., 2023).

* Corresponding author.
E-mail address: brita.elvevag@uit.no (B. Elvevåg).

https://doi.org/10.1016/j.psychres.2024.116119
Received 19 April 2024; Received in revised form 25 July 2024; Accepted 30 July 2024
Available online 3 August 2024

Accordingly, NLP methods have been heralded as promising instruments that could aid clinicians in the assessment and monitoring of thought disorders (Corcoran and Cecchi, 2020; Low et al., 2020; Voleti et al., 2023). However, several studies have identified numerous challenges with NLP-based assessments of thought disorder, such as a lack of generalizability across samples and languages (Parola et al., 2023) and questionable construct validity of computational measures (Hitczenko et al., 2021). Our own approach to study design has been underpinned by three assumptions: First, we have increased the probability of the speech signal of interest (e.g., incoherence) appearing at every measurement time point by creating tasks that push participants to their cognitive limit at which point an assay may be said to be sensitive enough to elicit this cognitive vulnerability, namely thought disorder. Second, these resulting script-like speech responses can thus be rated by expert humans on specific dimensions (e.g., amount of content remembered and overall coherence) and these ratings thus used to train machines to automate the process and optimize design for longitudinal studies (Chandler et al., 2020a, 2021a). Third, since these speech prompts very likely elicit the signals of interest, and can be fully automated (e.g., for monitoring purposes), crucially their very design enables the necessary eventual safeguarding against dangerous machine learning errors as they are well suited to build in methods for detecting when a model is out of bounds, which is absolutely essential for any automated system (e.g., Chandler et al., 2021b).

In an early study on NLP applications in psychiatry, Elvevåg et al. (2007) used one of the first word-embedding based vector models, Latent Semantic Analysis (LSA), to detect thought disorder from speech and predict human ratings (e.g., coherence) assigned to speech samples. This work introduced a novel methodology: by representing words or sentences as vectors in a high-dimensional semantic space, concepts such as semantic incoherence or tangentiality could be computationally instantiated by comparing the similarity of vectors across a spoken response. Subsequently, LSA was further employed to detect subtle deviations in discourse between patients with schizophrenia, their-first degree relatives, and unrelated healthy controls (Elvevåg et al., 2010; Rosenstein et al., 2015). This methodology was later applied to predict the onset of psychosis in high-risk populations (Bedi et al., 2015) where it was shown to be effective across different speech collection protocols and risk cohorts (Corcoran et al., 2018). As word- and sentence-embedding methods continued to evolve, more sophisticated embedding methods (e.g., GloVe, word2vec) outperformed LSA in detecting incoherence from speech (Iter et al., 2018). Other computational methods, such as speech graph connectivity (Mota et al., 2012), measures of referential cohesion (Gupta et al., 2018; Iter et al., 2018) and measures of language connectedness (Voppel et al., 2021) have also been developed to analyze language and identify indicators of thought disorder. Notably, these computational approaches have shown sensitivity in discriminating between diagnostic groups (e.g., schizophrenia vs. mania) exceeding that of traditional psychiatric scales like the Brief Psychiatric Rating Scale (BPRS) and the Positive and Negative Syndrome Scale (PANSS) (Mota et al., 2012), suggesting that these quantitative analyses are complementary, rather than redundant with existing psychometric scales.

In recent years, advances in deep learning and NLP have led to the development of powerful new neural network-based models. These models (e.g., BERT, RoBERTa) often called Large Language Models (LLMs) make use of the Transformer architecture (Vaswani et al., 2017) and extensive pre-training on large quantities of text, resulting in state-of-the-art performance on a variety of NLP tasks, including word and sentence embedding and classification. Accordingly, several recent studies have applied modern neural network-based sentence embedding methods to detect thought disorder in speech. For instance, Sarzynska-Wawer et al. (2021) used Embeddings from Language Models (ELMo) to represent interview responses (conducted in Polish) and found that ELMo embeddings were more accurate than word2vec-based coherence metrics in differentiating patients from healthy controls. In another study, Tang et al. (2021) used a BERT model to compute embeddings of both an interviewer prompt and participants' responses. Using these embeddings, they found that patients with schizophrenia showed higher levels of tangentiality or derailment than healthy controls, as quantified by the change over time in the semantic similarity of response sentences to the original prompt.

An even more recent advancement in NLP is the rise of generative LLMs (e.g., ChatGPT, Gemini, LLaMA). Generative LLMs represent a fundamentally new paradigm in NLP, which has been called "prompt-based learning" (Liu et al., 2023). Unlike traditional supervised learning, which depends on labeled training data to build a predictive model, generative LLMs can be "prompted" with written instructions in natural language to elicit predictions or generate text. They have also been accompanied by web-based user interfaces and APIs (application programming interfaces; software interfaces that allow programmatic access to the models) which have made these methods accessible to a much wider audience of users. Accordingly, the field of psychiatry has investigated the use of generative LLMs for a variety of applications, including answering clinical questions related to psychiatric diagnosis and treatment (Luykx et al., 2023), responding to common patient questions in mental healthcare (Grabb, 2023), automated textual analysis for personality estimation (Amin et al., 2023), suicidal tendency detection (Amin et al., 2023; Lamichhane, 2023; Xu et al., 2023), depression detection (Lamichhane, 2023; Sadeghi et al., 2023; Xu et al., 2023; Yang et al., 2023) and sentiment analysis (Rathje et al., 2023). LLMs have also received increased interest and investment in the broader field of healthcare (see Thirunavukarasu et al. (2023) for a recent review of LLMs in medicine in general). In one study highly relevant to the current investigation, researchers used GPT-4 to automatically generate fluency ratings for spoken responses to a picture description task (e.g., evaluations of how fluently the participant described the picture, including identifying key elements of the picture, repetitiveness, and the presence of unclear phrases). Subsequently, they generated BERT embeddings to represent these GPT-4 fluency evaluations, which were used along with embeddings of the original description to classify which participants had a diagnosis of Alzheimer's Disease (Bang et al., 2024).

Despite the increased use of these models in the field, there has been very little research investigating the consistency and reliability of their output. As stated in a recent intergovernmental organization's report on responsible AI for health, "while generative AI is powerful, its creative nature increases uncertainty and thereby risk for the delivery of healthcare" (Anderson and Sutherland, 2024, p. 11). Unlike prior NLP methods applied to measuring mental states, generative LLMs are stochastic models (e.g., non-deterministic). Put differently, they may generate completely different content, ratings, or decisions given the same input. Thus, the importance of assessing the consistency of a model's output and understanding potential sources of variability cannot be overstated. It is a key factor in establishing trust in AI systems, a major barrier to the development of real-world applications for mental healthcare (Chandler et al., 2020b). In addition to eroding trust by clinicians, patients, or other stakeholders, inconsistent outputs could lead to misdiagnoses, inappropriate treatment recommendations, or other potentially harmful consequences in a clinical setting.

In this paper, we have addressed this research gap by using modern LLMs to re-analyze a subset of the data from Elvevåg et al. (2007), collected in a structured interview designed to elicit uninterrupted speech from participants (Experiment 4 in Elvevåg et al. (2007)). Our study aimed to assess the ability of generative LLMs (specifically, the latest models in the OpenAI GPT series and the open source Llama 3 model) to predict human ratings of coherence, content, and tangentiality assigned to the spoken responses, and to compare these LLMs with methods used in previous research. Furthermore, we evaluated the consistency of LLM-generated ratings. Put differently, we investigated the degree to which LLMs generated the same ratings given the same responses over subsequent trials, and the effect of consistency on prediction accuracy. Through this investigation, we aimed to understand

what factors influenced the consistency of LLM output, recognize limitations in these models, and identify best practices for ensuring the most accurate and reliable output when utilizing these powerful models for psychiatric textual analysis.

1.1. Participants

Participants included 26 patients with a formal diagnosis of schizophrenia and 25 healthy controls, providing a diverse sample with which to conduct our analysis. Patients met the DSM-IV criteria for schizophrenia or schizoaffective disorder, as determined using the Structured Clinical Interview for DSM-IV (SCID), with three psychiatrists reaching a consensus diagnosis. Further, the severity of thought disorder was assessed in patients using a standard clinical measure (the Scale for the Assessment of Thought, Language and Communication; TLC (Andreasen, 1986)), as described in Elvevåg et al. (2007). Notably, these TLC ratings were assigned based on a clinical interview separate from the speech tasks analyzed in this study. Patients' global TLC scores (on a scale of 0 to 4, with higher scores representing greater severity of thought disorder) ranged from 0 (absent) to 3.8 (mean = 1.81, SD = 1.10), indicating a wide range of severity among the patients in this sample. Characteristics of the patient and control samples are shown below in Table 1.

1.2. Dataset

The dataset analyzed in this study is a subset of data from Elvevåg et al. (2007), specifically focusing on the spoken (manually transcribed) responses to two questions that were asked during the course of a structured interview conducted in English: (1) *"Could you tell me the story of Cinderella?"* and (2) *"What are the steps involved when people have to do laundry?"* These tasks were designed to elicit sufficient speech from participants in a systematic manner to allow reliable human assessment of dimensions such as coherence, tangentiality and content. Participants were encouraged to talk for as long as they wished, and there was no time limit. The dataset contained responses to both questions from all participants in our sample, with the exception of one control participant who did not have a response for the Cinderella task, although their data was not excluded from analysis on the laundry task. All transcripts were manually checked to verify that they did not contain any personally identifiable information. In addition to the transcripts of each response, the dataset also contained human-annotated ratings of coherence, content, and tangentiality for each response. Each of these three dimensions was blindly rated by two experts (i.e., the experts were blind to diagnostic group membership, but not to the objectives of the study), using a scoring rubric developed in Elvevåg et al. (2007) (see supplementary material), in order to characterize the type and level of thought disorder present in each response. The consistency among the two raters was good (intra-class coefficients were 0.85, 0.97 and 0.94 for coherence, content and tangentiality, respectively). Since agreement was high, and to maintain the independence of sample ratings for analysis, the ratings were averaged across the two raters to generate a single ground truth human rating for each response. Table 2 shows the questions used in both tasks, along with sample responses (one from each task for both patient and control participants) and their corresponding human ratings.

Table 1
Characteristics of patient and control participants in the sample.

	Patients (n = 26) 19 M, 7F		**Controls (n = 25)** 10 M, 15F	
	Mean	SD	Mean	SD
Age in years	33.8	7.6	35.4	12.9
WAIS-R IQ*	94.5	13.4	108.0	11.9
WRAT-R IQ*	103.6	11.5	109.3	8.7
Age at 1st hospitalization (years)	21.5	4.2	N/A	
TLC global score (0–4 scale)	1.8 (Range: 0–3.8)	1.1	N/A	

Intellectual function was assessed with the Wide Range Achievement Test-Revised Reading (WRAT-R; Jastak and Wilkinson, 1984) and a short form of the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981).
*$p < .05$ (independent samples t-test).

As a preliminary step, we investigated the relationship between the scores from traditional ratings of patients' thought disorder (i.e., global TLC scores) and the three human ratings of the spoken responses to the Cinderella and laundry prompts, which were found to be positively correlated (Pearson R values range from 0.17 to 0.64), although only three of the correlations were statistically significant (Table 3). We also computed the correlations between the three human ratings within each task. The results (Table 3) show that the three ratings were moderately correlated with each other (Pearson R values range from 0.46 to 0.74), although the magnitudes of the correlation are low enough to suggest the ratings capture distinct dimensions of thought disorder. Importantly, using human ratings of these responses allows us to move beyond group differences (e.g., classifying patients vs. control participants, which themselves are heterogeneous groups that may contain diverse presentations of the symptoms of interest), and focus on the actual disturbances in speech that are known to be present in the data (Hitczenko et al., 2021).

1.3. Research questions and experiments

Two experiments were designed to evaluate the accuracy and consistency of state-of-the-art LLMs in predicting human ratings (labels) of coherence, content, and tangentiality from the transcripts in this dataset. We also compared the performance of LLMs with more traditional supervised machine learning models using NLP features that have been applied in previous research. The experiments were guided by the following research questions.

Question 1: How accurately can state-of-the-art LLMs predict the human ratings of coherence, content, and tangentiality compared to a traditional supervised machine learning model that uses NLP features from prior research?

Question 2: What effect do different model versions (e.g., GPT-3.5 vs. GPT-4 vs. Llama 3) and parameters (e.g., temperature values; see Section 2.1.3) have on LLM predictions?

Question 3: How consistent are LLM predictions across multiple trials in this rating task (e.g., how much variance is there when a given response is rated multiple times)?

Experiment 1 addressed the first and second questions by (1) developing supervised baseline models that used NLP features from previous research to predict the three human ratings, and then (2) comparing the accuracy of these baseline models' predictions with ratings predicted by the LLM, in order to determine if LLMs can improve on the performance of the previous generation of models. Further, Experiment 1 investigated different LLM versions and parameters, and the effect they had on rating predictions. Finally, Experiment 2 answered the third question by quantifying the consistency of LLM-generated ratings across multiple trials.

2. Experiment 1: Accuracy of model ratings

2.1. Methods

2.1.1. Training of supervised baseline models

In order to determine how well LLMs may perform relative to traditional NLP models, we first created a baseline model for each label (coherence, content, tangentiality), and their rating accuracies were compared to the accuracies of the LLMs. Our goal was to generate a reasonable benchmark of traditional NLP performance with which to compare our results. Because the three labels differ in the underlying constructs that they each measure (see Table 3), we developed three

Table 2
The two speech tasks analyzed in this study, along with a sample response to each (for both patient and control participants) and the corresponding human ratings assigned to the sample responses. Note: the sample responses shown below are based on actual responses in the dataset but are not presented verbatim and have been lightly edited as an extra precaution to preserve anonymity.

Prompt	**Could you tell me the story of Cinderella?**	**What are the steps involved when people have to do laundry?**
Patients	$N = 26$ responses	$N = 26$ responses
Mean number of words (range)	237 (34–634)	139 (35–684)
Sample response	"um let's see cinderella she uh I forget how it starts she gets hooked up with her dad hooks up with the evil stepmother who has evil kids and they make her do the housework and chores then cinderella meets a fairy godmother who can make her wishes come true so she goes to the ball and meets the prince loses a slipper they get married or something"	"well you're living with people so you have to have your own day when you do your laundry a day when you get all your dirty clothes you put them in either hot or cold water hot is for the same colors like white if you put a whole bunch of stuff in there the same color or different colors you put it on cold and you uh cold and you set it on large and when it's done you put it in the washing machine or in the dryer I mean and that's it you're done so you put your clothes away and you have nice clean clothes"
Human ratings for sample response (1–7 scale)	Coherence: 3 Content: 2.5 Tangentiality: 1	Coherence: 3.5 Content: 4 Tangentiality: 2.5
Controls	$N = 24$ responses	$N = 25$ responses
Mean number of words (range)	452 (170–1292)	162 (36–603)
Sample response	"cinderella was the um the stepdaughter of somebody who got stuck in the house doing all the chores and um the uh daughters of the mother of the house were invited to the ball and cinderella wanted to go to the ball but she wasn't able to go and then she found her fairy god mother who made her a nice dress and gave her a um transportation but told her she to be back by midnight so she went to the ball and it was a smashing success um but it was close to midnight and she was dancing with the prince who liked her and she ran out at the stroke of midnight and lost her shoe her uh let's see then he went around looking for whoever belonged to the shoe and it fit cinderella and they got married and lived happily ever after"	"well first you need dirty clothes to wash but generally speaking you would uh sort the clothes um by colors so separate those for soak and general soak um put 'em in the washing machine um add the cleaning ingredients put the top down turn the knobs and put it on to start um then wait until it's um done pull 'em out put 'em in the dryer and dry 'em out"
Human ratings for sample response (1–7 scale)	Coherence: 1.5 Content: 2.5 Tangentiality: 1	Coherence: 1 Content: 2 Tangentiality: 1

Table 3
First row: Correlations (Pearson R) between global TLC scores and the three human ratings on both the Cinderella and Laundry speech tasks, for patients only (as control participants were not given a TLC interview). Below: Within-task correlations between the three human ratings, computed using all data in the sample (both patient and control participants).

	Cinderella			**Laundry**		
	Coh.	**Cont.**	**Tang.**	**Coh.**	**Cont.**	**Tang.**
TLC Score (Patients)	0.59*	0.37	0.36	0.64*	0.54*	0.17
Coherence	–	0.58*	0.55*	–	0.74*	0.64*
Content	–	–	0.46*	–	–	0.49*
Tangentiality	–	–	–	–	–	–

Note: Coh. = Coherence, Cont. = Content, Tang. = Tangentiality.
*$p < .01$

distinct sets of language features for our baseline models. The feature sets were based on previous research (Elvevåg et al., 2007; Iter et al., 2018; Morgan et al., 2021; Tang et al., 2021), and the technical details of how the feature sets were derived for each of the three labels are presented in the supplementary material. After extracting these feature sets for the three labels, we trained Random Forest Regressor models to predict the human ratings of these labels. We used a leave-one-out cross validation scheme to generate predictions for each transcript and evaluate the accuracy of these predicted ratings.

2.1.2. Prompt design for LLM ratings

Next, we developed a prompt to elicit ratings for each transcript from the LLM. Unlike traditional supervised NLP approaches which aim to learn predictive patterns from labeled data, LLMs can be "prompted" to generate predictions without using specific examples as training data. However, this paradigm, known as "zero-shot learning" (Kojima et al., 2022) depends heavily on the quality of the prompt used to query the LLM (Liu et al., 2023), and as such, developing good prompts that elicit accurate predictions for a given task is an important part of the process (Wang et al., 2023). In this study, we used a prompt similar to the one outlined in Naismith et al. (2023). Although this example comes from a different domain (second language learners), the application is very similar to ours, as they demonstrated promising performance in using GPT-4 to rate written discourse coherence in a manner consistent with expert human raters. Our prompt contained four sections: (1) TASK: a detailed description of the task to be completed by the LLM and desired output, (2) RUBRIC: the rubric by which to rate the responses (the same rubric used by the human raters in the original dataset - see supplementary material), (3) GUIDELINES: a more specific set of instructions for the LLM to complete the task and generate output in the desired format, and (4) TEXT: the transcript of the question, and the corresponding response to be rated. An example of a full prompt for the coherence label on the Cinderella task is shown in Fig. 1.

2.1.3. Generation of LLM ratings

We used the prompt shown above to elicit ratings from the LLMs. Specifically, we used GPT-3.5, GPT-4, and Llama 3. Our general procedure is illustrated below in Fig. 2. For the GPT models, we queried the model using the chat completions endpoint of the OpenAI API (*OpenAI Platform*, 2023). In this experiment, we tested two models "gpt-3.5-turbo" and "gpt-4″ as these were the two newest models at the time of initial testing (October 2023). Although we expected GPT-4 to perform better, we wanted to compare its performance with GPT-3.5, as it is more widely accessible than GPT-4 (due to usage limits and significantly higher cost (80 times) of GPT-4). Additionally, in order to compare the GPT models with an open source alternative, we utilized the Meta Llama 3 model (AI@Meta, 2024). Specifically, we queried the "meta-llama-3–70b-instruct" model using the replicate.com API. In addition to comparing these three LLMs, we also tested different values

```
# TASK

You are rating the coherence of a piece of text. The text is a transcript of a spoken response to the question: "Co
uld you tell me the story of Cinderella?".
Provide a numeric rating of the coherence of this response on a Likert scale of 1-7 (1 being the most coherent, 7 b
eing the least coherent). Ratings at intervals of .5 (e.g., 3.5) are permitted. Additionally, generate an explanati
on for why the given rating was selected. Base the rating and explanation on the rubrics and guidelines below. Resp
ond in JSON format, with one field for the coherence rating, and one for the explanation (e.g., {"Coherence Ratin
g":rating, "Coherence Rating Explanation":explanation}). Do not include any additional text except for the JSON des
cribed.

# Rubric

The rating scale is from 1 to 7, with ratings at intervals of .5 also permitted. The following benchmarks describe
features of the text that are indicative of the given rating:
1 - Passage makes sense, words are ordered.
2.5 - Some instances of mild disorganization, but can be interpreted. Filler, ambiguous referents, or mild instance
s of sequencing problems (in "process" time) may be present.
4 - Passage has multiple instances of between clause or within clause derailment / loose associations, but meaning
can be discerned. Failures in connectivity or poverty of content. Many instances out of sequence.
5.5 - Much of passage is incomprehensible.
7 - Whole passage is incoherent due to clause failures, lacking connection. "word salad".

# GUIDELINES

* You must choose one rating only
* Provide specific examples to support the rating and explanation given
* Use language from the response and the rubric to justify and explain the rating
* Write in the third person
* Remember that the text is a transcript of a spoken response, and spoken language differs significantly from writt
en text
* For borderline cases, say why the rating is not higher or lower

# TEXT

Question: "Could you tell me the story of Cinderella?"
Response: "[Insert transcript of response here]"
```

Fig. 1. Example of a prompt used to query the LLMs. This specific prompt is for the coherence label and the Cinderella prompt.

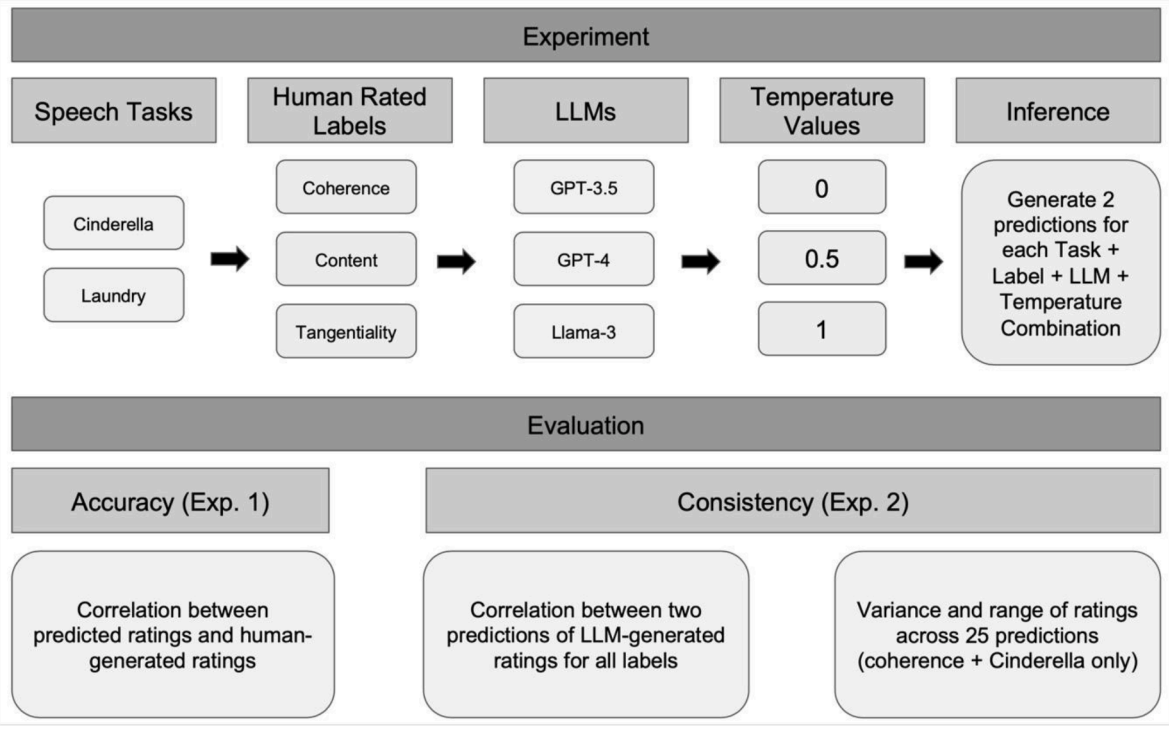

Fig. 2. The procedure for eliciting ratings from the LLMs and evaluating accuracy and consistency.

for the "temperature" parameter to investigate its effect on model accuracy and consistency. The temperature parameter is described in the OpenAI API documentation (as of October 2023): *"The API is non-deterministic by default. This means that you might get a slightly different completion every time you call it, even if your prompt stays the same. Setting temperature to 0 will make the outputs mostly deterministic, but a small amount of variability will remain. […] Lower values for temperature result in more consistent outputs, while higher values generate more diverse*

and creative results." The temperature parameter can range from 0 to 2 for the OpenAI models, and 0 to 5 for Llama 3 (via replicate.com). For all three LLMs, we tested temperature values of 0, 0.5, and 1 (as we found that the models would not reliably generate output in the desired format using values greater than 1).

We generated two predictions for each label (coherence, content, and tangentiality) with these three temperature values for each transcript from the Cinderella and laundry tasks using the GPT-3.5, GPT-4, and Llama 3 models. To assess model accuracy, we computed a Pearson correlation to the human ratings (averaged across the two generated predictions).

2.2. Results

2.2.1. Accuracy of supervised baseline models

Pearson correlations between the baseline model-predicted ratings and human ratings are shown below in Table 4. Overall, the correlations were low to moderate, and performance was worse on the laundry task than Cinderella. Notably, at the level of individual labels, our baseline predictions were less accurate than every LLM-generated prediction with the exception of tangentiality on the Cinderella task, where the baseline model yielded the best performance (correlation of 0.60).

2.2.2. Accuracy of LLMs

The Pearson correlations between the LLM-generated ratings and the human ratings are also shown in Table 4. Since two iterations of ratings were generated (to assess model consistency, see Experiment 2), we computed the correlations to human ratings separately for each iteration, and we report the mean in Table 4. Our results indicate that all three LLMs were able to predict human ratings for coherence, content, and tangentiality with a moderate degree of accuracy. On average, GPT-4 matched or exceeded GPT-3.5 performance in both the Cinderella (highest mean correlation of 0.49 vs. 0.48) and Laundry (0.66 vs. 0.58) tasks, while Llama 3 matched GPT-4 performance in the Cinderella task (also achieving 0.49) and outperformed GPT-4 in the laundry task (0.69 vs. 0.66).

These results indicate that on average, LLM-generated predictions are far more accurate to human ratings than the supervised baseline predictions. This demonstrates that the "zero-shot" method used to elicit ratings from LLMs is a promising approach to predicting human ratings of thought disorder. It also suggests that LLMs are able to use the rubric included in the prompt to predict these ratings in a manner consistent with humans, even without access to labeled training examples. However, it also highlights some limitations to our baseline method. One possible explanation for the poor performance of the baseline models is that given the small size of our dataset, the supervised models did not have enough data to learn patterns from the baseline feature sets that reliably predict the human ratings. Another reason for the low performance could be that the feature sets we chose were suboptimal, and accuracy could be improved by modifying these features or adding additional features. However, in this paper we chose not to conduct additional experiments to try to improve our supervised baseline results, because (1) altering our features after evaluating their performance on the dataset (e.g., after performing cross-validation) may violate the independence of training and test data and lead to overly optimistic results, and (2) the primary goal of our paper is to evaluate the accuracy and consistency of LLM-generated predictions, not to conduct a comprehensive comparison with traditional supervised NLP methods. Thus, for the rest of the paper we focus on the factors affecting the consistency of LLM-generated ratings.

3. Experiment 2: Consistency of model ratings

3.1. Methods

3.1.1. Consistency across two iterations

This experiment aimed to assess the consistency of LLM ratings across multiple trials and investigate the factors that influence this consistency. The procedure used to generate ratings for each transcript with different model and temperature combinations is described above in Experiment 1 (see Fig. 2). Critically for this experiment, we generated two iterations of predictions for each transcript, model, and temperature combination. Essentially, generating two iterations is analagous to having human raters rate the same transcripts twice without remembering how they rated it the first time. Having two iterations of ratings allows us to assess the consistency of LLM predictions by investigating the extent to which the ratings are correlated across trials (e.g., a highly consistent model would yield a high correlation across the two trials). To do so, we computed a Pearson correlation between the two iterations of ratings for each model and temperature combination. This design also allows us to determine the effect that the temperature parameter has on model consistency and identify possible tradeoffs between accuracy and consistency.

3.1.2. Consistency across 25 iterations

In addition to evaluating consistency using the two iterations of ratings as described above, we also sought to quantify the consistency of ratings across a larger number of iterations ($n = 25$). To do so, we focused specifically on the Cinderella task and the coherence label. This decision was made because the cost of running a large number of iterations for all three labels on both tasks was prohibitive, so we selected a single task and label to conduct experiments with a higher number of iterations. The same methods described in Experiment 1 were used to generate an additional 25 iterations of ratings for all three models (GPT-3.5, GPT-4, and Llama 3), and all three temperature values (0, 0.5, 1). These experiments were conducted with the GPT models approximately

Table 4
Accuracy of Baseline and LLM Ratings. Table shows Pearson correlations between LLM-generated ratings (mean of two iterations) and human ratings for all three models and temperature parameter values (0, .5, and 1). Note: Coh. = Coherence, Cont. = Content, and Tang. = Tangentiality.

Task	Cinderella				Laundry			
Label	**Coh.**	**Cont.**	**Tang.**	**Mean**	**Coh.**	**Cont.**	**Tang.**	**Mean**
Baseline	0.22	0.31	0.60	0.38	0.05	0.18	0.25	0.16
GPT-3.5								
Temp = 0	0.45	0.49	0.49	0.48	0.56	0.58	0.59	0.58
0.5	0.54	0.46	0.28	0.43	0.39	0.52	0.46	0.46
1	0.48	0.40	0.23	0.37	0.40	0.47	0.52	0.46
GPT-4								
0	0.59	0.39	0.46	0.48	0.62	0.65	0.65	0.64
0.5	0.54	0.45	0.43	0.47	0.63	0.61	0.73	0.66
1	0.54	0.47	0.46	0.49	0.69	0.56	0.65	0.63
Llama 3								
0	0.57	0.39	0.46	0.47	0.69	0.66	0.71	0.69
0.5	0.56	0.44	0.48	0.49	0.68	0.67	0.73	0.69
1	0.57	0.37	0.44	0.46	0.66	0.68	0.69	0.68

4 months after our initial testing (February 2024), and with the Llama 3 model in July 2024.

Using these new distributions of 25 ratings per response, we quantified the consistency in three ways. First, for each of the $n = 50$ responses (e.g., 1 response for each of the 50 participants), we computed the variance of the predicted ratings across all 25 iterations. This metric allowed us to quantify the spread of the distribution of ratings for each individual response. Second, for each response we computed the range of the predicted ratings across the 25 iterations. This metric quantified the difference between the most extreme ratings (out of 25) for a given individual's response. Thus, while the variance provided an estimate of how much (on average) a single rating differed from the mean, the range gave a worst-case estimate of how far apart two ratings for the same response could be. For our third metric we computed the overall accuracy for each of the 25 iterations using the same methods described above (Pearson correlation of predicted ratings to human ratings). This metric allowed us to visualize the overall distribution of accuracy across the 25 iterations, as well as capture informative statistics on the range of accuracies (e.g., mean, max, and min correlations).

Finally, we used the distributions of 25 ratings to investigate a technique for improving the accuracy and stability of our predictions: ensemble learning. An ensemble prediction system works by combining the predictions of multiple models, which can potentially result in superior predictive performance than a single model alone (Nori et al., 2023). In this experiment, we created a simple ensemble prediction by computing the mean predicted rating for each response across all 25 iterations. The goal of this experiment was to compare the accuracy of the ensemble prediction with the overall distribution of accuracies from the 25 iterations.

3.2. Results

3.2.1. Consistency across two iterations

Pearson correlations between the two iterations of LLM-generated ratings are shown below in Table 5. Our findings indicate that the consistency of LLM ratings is highly dependent on both the model choice and the temperature parameter. Notably, Llama 3 exhibited the highest consistency of the three models across all temperature values (mean correlations ranged from 0.94 to 1.00). At temperature = 0, Llama 3 showed perfect consistency (correlations of 1.00) for all labels and maintained high consistency (mean correlation >= 0.94) even at temperatures of 0.5 and 1. Similarly, both GPT-3.5 and GPT-4 exhibited high consistency at temperature = 0, with mean correlations ranging from 0.96 to 0.98. However, we found that at higher temperature values of 0.5 and 1 there were notable differences between the consistency of GPT-3.5 and GPT-4. For example, at a temperature of 0.5, GPT-3.5 showed only moderate consistency (mean correlations of 0.70 and 0.62 in the Cinderella and laundry tasks, respectively) compared to GPT-4, which remained highly consistent (mean correlations of 0.92 and 0.95, respectively). This gap in consistency widened even further at the higher temperature value of 1.

The overall trend shows that the consistency of GPT-3.5 is highly impacted by the temperature parameter, GPT-4 remains relatively consistent even at higher temperature values, and Llama 3 maintains very high consistency across all temperature values tested. This suggests that the selection of a lower temperature value is crucial to generating consistent output, particularly with GPT-3.5. This finding is concerning, given that many studies which have used GPT models for analysis did not report a selection of the temperature parameter at all. Additionally, the default value for this parameter in the API and web interface is 1, likely because this value is optimized for the generation of creative content (which may be the primary use case for many users). However, this suggests that by default, much of the output from these models would be highly inconsistent across multiple trials.

3.2.2. Consistency of coherence ratings across 25 iterations

The distributions of the three metrics used to quantify consistency across 25 iterations are shown below in Fig. 3. Kernel Density Estimate (KDE) plots were created to visualize the distributions. In addition to the plots shown in Fig. 3, complete descriptive statistics (including the mean, max, and min correlations across the 25 iterations, as well as the mean variance and mean range) for these results can be found in the supplementary material.

The large sample of iterations ($n = 25$) in this experiment provided additional insight into the consistency of LLM-generated ratings. In line with our previous findings (Table 5), GPT-3.5 showed much higher variance than GPT-4 at higher temperatures, while having slightly lower variance at temperature = 0. Similarly, Llama 3 showed by far the lowest variance across all temperature values, producing a mean variance of 0.056 even at temperature = 1, compared to 0.511 and 0.194 for GPT-3.5 and GPT-4, respectively (see supplementary material). A similar pattern was observed in the range of ratings across 25 iterations. Notably, at temperature = 1, the mean within-response range for GPT-3.5 was 2.38, compared to 1.21 for GPT-4, and only 0.54 for Llama 3. This finding illustrates that Llama 3 maintains very high consistency even at higher temperature values, as on average, any two individual ratings (out of 25 trials) for the same response only differed by at most half a point (on the 7-point scale). Overall, these results illustrate that given an appropriate choice of model and temperature value (e.g., Llama 3, or GPT-4 with temperature = 0.5), the ratings can be quite consistent over 25 iterations. However, at higher temperature values, particularly with the older GPT-3.5 model, the consistency can decrease substantially.

The third row of Fig. 3 illustrates the effect of inconsistency on the overall accuracy of the ratings. For all three models, the range of accuracies (Pearson R values) increases with the temperature value. In

Table 5
Consistency of LLM Ratings. Table shows Pearson correlations between two iterations of LLM-generated ratings, for all three models and temperature values. Note: Coh. = Coherence, Cont. = Content, and Tang. = Tangentiality.

Task	Cinderella				Laundry			
Label	**Coh.**	**Cont.**	**Tang.**	**Mean**	**Coh.**	**Cont.**	**Tang.**	**Mean**
GPT-3.5								
Temp = 0	0.97	0.97	1.00	0.98	1.00	0.97	0.91	0.96
0.5	0.70	0.81	0.58	0.70	0.59	0.86	0.39	0.62
1	0.36	0.48	0.11	0.32	0.18	0.53	0.26	0.32
GPT-4								
0	0.93	0.96	0.99	0.96	0.96	0.98	0.96	0.97
0.5	0.93	0.95	0.89	0.92	0.98	0.94	0.92	0.95
1	0.71	0.90	0.85	0.82	0.81	0.88	0.96	0.88
Llama 3								
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.5	0.97	0.96	0.95	0.96	0.99	0.98	0.98	0.98
1	0.95	0.95	0.91	0.94	0.98	0.98	0.97	0.98

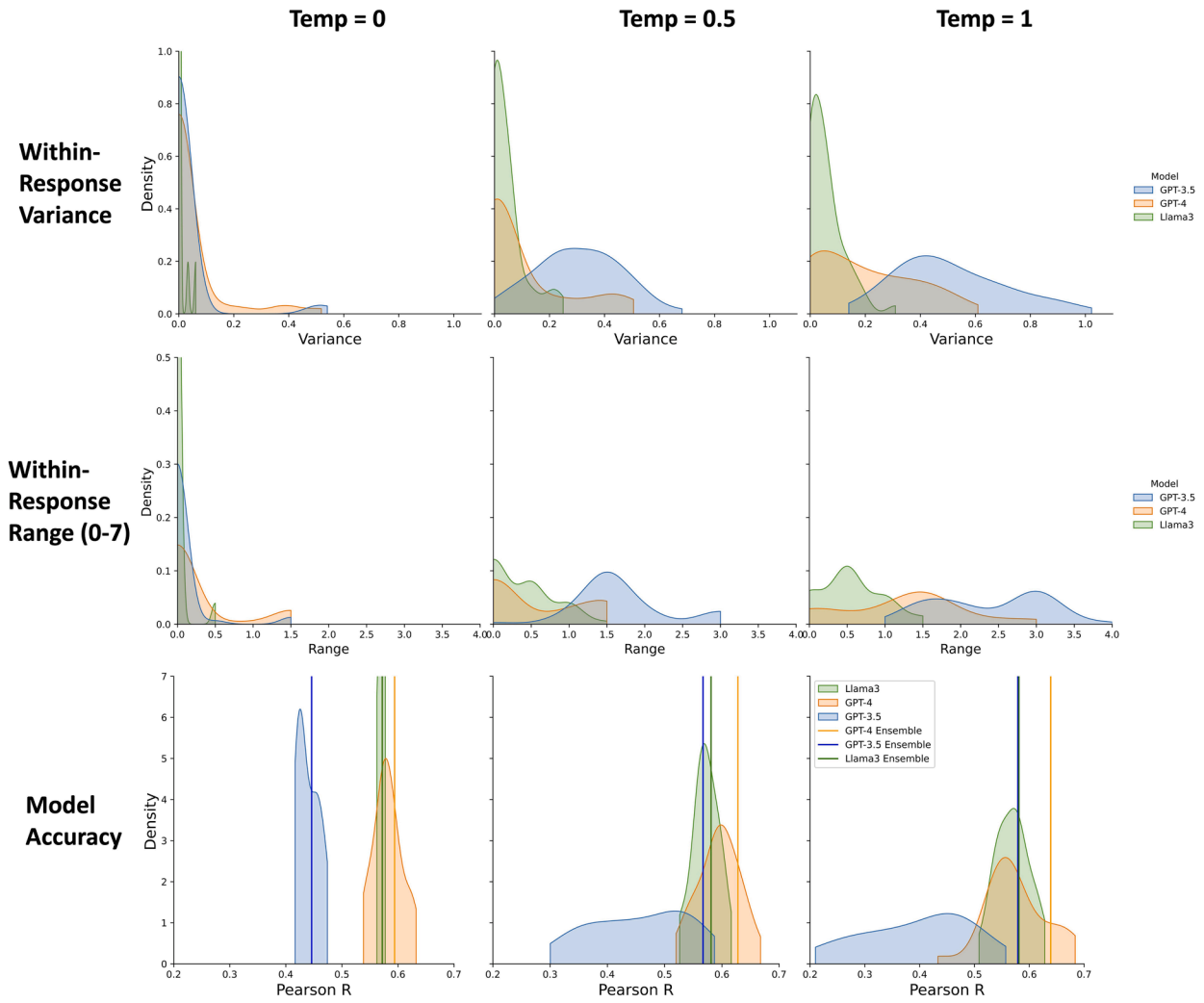

Fig. 3. Consistency metrics using 25 iterations of ratings for the coherence label (Cinderella task). The first row shows the distributions of within-response variance (computed across 25 ratings for all 50 responses). The second row shows the distributions of the range of ratings for each individual response (also computed across 25 ratings for all 50 responses), and the third row shows the distribution of overall accuracies (Pearson R correlations with human labels) for all 25 iterations. The vertical lines in the third row show the accuracy of the ensemble predictions described above.

other words, a higher temperature value will yield a wider distribution of accuracies, particularly with GPT-3.5. However, for all three models the highest mean correlation (over 25 iterations) was obtained with temperature = 0.5 (see supplementary material), suggesting an advantage in terms of accuracy for intermediate temperature values, albeit at the cost of greater variance.

The ensemble method used to aggregate predictions across iterations proved to be an effective way to improve accuracy, particularly at higher temperature values in models with higher variance. As seen in Fig. 3, for all three models, the accuracy of the ensemble prediction (colored vertical lines) was higher than the mean accuracy, and at the higher temperatures this was especially evident. The Llama 3 ensemble prediction improved performance the least (relative to the GPT models), likely because of the low levels of variance in the Llama 3 ratings to begin with. Still, the Llama 3 ensemble method was more accurate than 60–68% of the 25 individual iterations, depending on temperature value (see supplementary material). The accuracy of the GPT-3.5 ensemble prediction was higher than 96% of the 25 individual iterations at temperature = 0.5, and higher than 100% of iterations at temperature = 1.0, suggesting a significant improvement in accuracy from ensembling. Likewise, for GPT-4, the ensemble prediction was more accurate than 80 and 84% of the 25 iterations at temperature = 0.5 and 1.0, respectively, and the highest overall accuracy (Pearson R = 0.639) was obtained from the GPT-4 ensemble with temperature = 1. This finding indicates that there is an advantage to using a higher temperature if predictions are ensembled across multiple iterations and suggests that the higher variability in ratings leads to a more accurate ensemble prediction. This also shows that ensembling predictions can be an effective way to mitigate the inconsistency inherent in LLM output, as it was shown here to improve accuracy on average, and using the mean value across multiple iterations is robust to the variability of individual ratings.

4. Discussion

The results of our study provide valuable insights into the use of generative LLMs for predicting human ratings of thought disorder in the context of psychiatric textual analysis. Our findings demonstrate that LLMs, specifically GPT-3.5, GPT-4, and Llama 3 can generate ratings that correlate moderately well with human ratings. This suggests that LLMs have potential utility in psychiatric research settings, particularly for use in tasks that involve linguistic analysis and rating. However, the inconsistency of these models' output highlights potential limitations and steps needed to improve their reliability.

The observed difference in consistency between GPT-3.5, GPT-4, and Llama 3, especially at higher temperature values, emphasizes the need for careful model selection and parameter tuning when applying LLMs in

psychiatric research. High variability in LLM's output could potentially lead to inconsistent results, which might negatively impact decision-making processes in a clinical setting. Further, inconsistent output may undermine trust in AI-based systems, limiting the potential for these powerful tools to eventually be utilized in real-world settings. This is particularly concerning for the field because many recent studies have employed these GPT models without any mention of parameter selection (e.g., temperature value) or evaluation of output consistency across multiple trials. Our findings show that resorting to GPT's "default" parameters may result in inconsistent output and results that do not replicate. Our results also highlight several possible approaches for mitigating LLM inconsistency. We found that ratings were more consistent when using GPT-4 with lower temperature values (0 and 0.5), and that overall Llama 3 produced much more consistent ratings than the GPT models. Further, we found that model accuracy could be increased via a simple ensemble method, which improves consistency by averaging across multiple iterations of ratings. Notably, the ensemble method worked best with higher temperature values. This finding demonstrates an important tradeoff: higher temperature values will result in less consistent individual ratings, but a more accurate overall ensemble, while lower temperature values yield more consistent individual ratings yet a less accurate ensemble. Indeed, the ensemble method is analogous to having 25 separate raters pooling their judgments, and our results show that the more variable the raters' judgements are, the more accurate their pooled predictions will be.

4.1. Ethical considerations

This paper focused primarily on the issue of LLM consistency as a practical challenge to developing computational methods for assessing thought disorder. However, the application of LLMs in psychiatry also requires careful consideration of several ethical concerns (Diaz-Asper et al., 2024; Li et al., 2023). As discussed above, the risk of inconsistent assessments could lead to adverse consequences in a clinical setting (e.g., misdiagnosis) and a loss of trust by stakeholders, particularly given the lack of explainability of these models (He et al., 2023). Thus, it is essential to appropriately convey the limitations of these methods, ensure rigorous verification of their outputs, develop methods to increase their transparency, and design safeguards to manage these risks. Beyond the challenges associated with a lack of consistency, there are significant risks of LLMs producing biased or harmful content (Bender et al., 2021; Guo and Caliskan, 2021; Singhal et al., 2022), resulting in systems that may perpetuate bias pertaining to race, sex, or culture. Additionally, the use of LLMs in psychiatry presents a number of ethical concerns pertaining to data privacy (Minssen et al., 2020). Textual data input to LLM APIs may contain sensitive protected health information or personally identifiable information, and many commercial AI companies lack transparency in how they use or store this textual data (Li et al., 2023). Thus, another critical step is to implement strict controls for the de-identification of data that may be analyzed by an LLM. Alternatively, some models can be downloaded onto a secure and private server, which negates the risks of uploading sensitive information to a company's API, although thus far most research in the field has not used this approach. Obviously, much work remains to be done to address these challenges and work toward the responsible, trustworthy, and ethical application of AI in healthcare (Anderson and Sutherland, 2024; Diaz-Asper et al., 2024). This paper focused on one important piece of this work (evaluating LLMs for accuracy and consistency) and illustrates the critical need for developers and clinicians to understand and mitigate their limitations as they work to incorporate these powerful tools in psychiatric research and practice.

4.2. Limitations and future work

This study has a number of limitations, which highlight several directions for future research. First, our study only investigated three LLMs (GPT-3.5, GPT-4, and Llama 3), although there are many other models that could be compared (e.g., Gemini, Claude). We chose to focus on these GPT models because of their popularity (both with the general public as well as in recent research in psychiatry and medicine), and to highlight issues with their consistency. Further, we tested Llama 3 to include an open-source LLM to compare with the GPT models. However, future work should conduct a more comprehensive evaluation of the strengths and weaknesses of a larger group of LLMs in this domain (see Jin et al., 2023; Xu et al., 2023 as examples).

Second, we restricted our experiments to a "zero-shot" approach, where only a rubric and guidelines were provided to the LLM, as opposed to a "few-shot" approach where the model is also provided with a small number of labeled examples from the dataset. However, research in other domains has shown that providing even a few training examples can improve model performance (Wang et al., 2020), so future work can investigate this strategy and the effect it has on model accuracy and consistency. Similarly, we only investigated a simple ensemble strategy (averaging ratings across multiple iterations) for mitigating inconsistency. Future work should investigate additional prompting methods and ensemble strategies (e.g., chain-of-thought prompting, self-consistency prompting, ensemble refinement (see Singhal et al., 2023) for improving the consistency of LLM output in similar psychiatric applications).

Third, it is worth noting that this study analyzed speech that was transcribed manually, and thus does not represent a fully automated approach in its current form. However, this genre of speech (e.g., story recall) has been shown to be very suitable for automatic speech recognition (ASR), with fully automated systems yielding comparable results to manual transcription (Holmlund et al., 2020). Thus, this likely does not represent a significant barrier to implementing these methods as part of a fully automated, end-to-end system, although it would be necessary to have a method in place to check the accuracy of ASR transcription (Diaz-Asper et al., 2022). Also of note is that the speech data studied was in English only, which raises the question of how well the method generalizes to other languages. Existing LLMs do have multi-lingual capabilities, and other research has investigated features such as coherence in a variety of other languages (e.g., German, Danish, Chinese) with different levels of success, which may be attributed to factors in the model, or aspects of coherence in the language itself (Just et al., 2020; Parola et al., 2023). This is another important area for future research. Finally, we focused exclusively on the numeric ratings and did not analyze the explanations generated by the LLMs. We chose to prompt the model to generate explanations in addition to numeric ratings because previous work has shown that this method can produce more accurate results than generating a rating alone (Naismith et al., 2023). However, in this study, we report only on the accuracy and consistency of the numeric ratings, leaving further analysis of the explanations to future work. In subsequent research we aim to investigate both the quality and consistency of these explanations as well, as this is potentially a promising step towards developing more trustworthy and explainable AI systems.

4.3. Conclusion

The application of generative LLMs in psychiatry presents exciting opportunities, particularly in assessing thought disorders from speech. Our study has shown that these models are capable of predicting human ratings of thought disorder, even without task- or domain-specific tuning, or access to labeled training data. However, our study also identified major concerns with the inconsistency of LLM output. In light of these concerns, we showed that inconsistency could largely be mitigated by careful model and parameter selection. Further, our experiments indicated that LLM ensembles are a promising method to improve the consistency and reliability of LLM output, and we suggest that future research should investigate additional prompting and ensemble strategies to improve performance. We also recommend that future research

applying LLMs in psychiatric applications include rigorous evaluations of consistency, reporting and testing of model parameters, and focus on creating mechanisms to control and manage the inconsistency inherent in these models. In conclusion, the novel use of LLMs for assessing thought disorder, though promising, necessitates a thorough understanding of their functioning to ensure their eventual safe and effective application in clinical settings.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used GPT-4 in order to improve readability of some portions of the manuscript, as well as to create some python functions to aid in the analyses. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Samuel L. Pugh: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Chelsea Chandler:** Writing – review & editing, Software, Methodology. **Alex S. Cohen:** Writing – review & editing, Supervision. **Catherine Diaz-Asper:** Writing – review & editing, Supervision. **Brita Elvevåg:** Writing – review & editing, Supervision, Project administration, Data curation, Conceptualization. **Peter W. Foltz:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

A.S.C., has stock or related interests in Quantic Innovation which develops and validates digital health measurements. In the past three years, A.S.C has received honoraria/support from Indivior and Boehringer Ingelheim.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2024.116119.

References

AI@Meta. (2024). Llama 3 model card. Accessed June 27, 2024. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Amin, M.M., Cambria, E., Schuller, B.W., 2023. Will affective computing emerge from foundation models and general artificial intelligence? A First Evaluation of ChatGPT. IEEe Intell. Syst. 38 (2), 15–23. https://doi.org/10.1109/MIS.2023.3254179.

Anderson, B., Sutherland, E., 2024. Collective Action For Responsible AI in Health. OECD. https://doi.org/10.1787/f2050177-en.

Andreasen, N.C., 1986. Scale for the assessment of thought, language, and communication (TLC). Schizophr. Bull. 12 (3), 473–482. https://doi.org/10.1093/schbul/12.3.473.

Andreasen, N.C., Grove, W.M., 1986. Thought, language, and communication in schizophrenia: diagnosis and prognosis. Schizophr. Bull. 12 (3), 348–359. https://doi.org/10.1093/schbul/12.3.348.

Bang, J.-U., Han, S.-H., Kang, B.-O., 2024. Alzheimer's disease recognition from spontaneous speech using large language models. ETRI J 46, 96–105. https://doi.org/10.4218/etrij.2023-0356.

Bedi, G., Carrillo, F., Cecchi, G.A., Slezak, D.F., Sigman, M., Mota, N.B., Ribeiro, S., Javitt, D.C., Copelli, M., Corcoran, C.M., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ. Schizophr. 1, 15030. https://doi.org/10.1038/npjschz.2015.30.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: can language models be too big?. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. https://doi.org/10.1145/3442188.3445922.

Chandler, C., Foltz, P.W., Cohen, A.S., Holmlund, T.B., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Elvevåg, B., 2020a. Machine learning for ambulatory applications of neuropsychological testing. Intell. Based. Med. 1, 100006 https://doi.org/10.1016/j.ibmed.2020.100006.

Chandler, C., Foltz, P.W., Elvevåg, B., 2020b. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. Schizophr. Bull. 46 (1), 11–14. https://doi.org/10.1093/schbul/sbz105.

Chandler, C., Holmlund, T.B., Foltz, P.W., Cohen, A.S., Elvevåg, B., 2021a. Extending the usefulness of the verbal memory test: the promise of machine learning. Psychiatry Res. 297, 113743 https://doi.org/10.1016/j.psychres.2021.113743.

Chandler, C., Foltz, P.W., Cohen, A.S., Holmlund, T.B., Elvevåg, B., 2021b. Safeguarding against spurious AI-based predictions: the case of automated verbal memory assessment. In: Proceedings of the NAACL-HLT 2021 Workshop on Computational Linguistics and Clinical Psychology. https://www.aclweb.org/anthology/2021.clpsych-1.20.pdf.

Corcoran, C.M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D.C., Bearden, C.E., Cecchi, G.A., 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatr. 17 (1), 67–75. https://doi.org/10.1002/wps.20491.

Corcoran, C.M., Cecchi, G.A., 2020. Using language processing and speech analysis for the identification of psychosis and other disorders. Biol. Psychiatr. Cogn. Neurosci. Neuroimag. 5 (8), 770–779. https://doi.org/10.1016/j.bpsc.2020.06.004.

DeLisi, L.E., 2001. Speech disorder in schizophrenia: review of the literature and exploration of its relation to the uniquely human capacity for language. Schizophr. Bull. 27 (3), 481–496. https://doi.org/10.1093/oxfordjournals.schbul.a006889.

Diaz-Asper, M., Holmlund, T.B., Chandler, C., Diaz-Asper, C., Foltz, P.W., Cohen, A.S., Elvevåg, B., 2022. Using automated syllable counting to detect missing information in speech transcripts from clinical settings. Psychiatry Res. 315, 114712 https://doi.org/10.1016/j.psychres.2022.114712.

Diaz-Asper, C., Hauglid, M.K., Chandler, C., Cohen, A.S., Foltz, P.W., Elvevåg, B., 2024. A framework for language technologies in behavioral research and clinical applications: ethical challenges, implications, and solutions. Am. Psychol. 79 (1), 79–91. https://doi.org/10.1037/amp0001195.

Elvevåg, B., Foltz, P.W., Rosenstein, M., DeLisi, L.E., 2010. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. J. Neurolinguistics. 23 (3), 270–284. https://doi.org/10.1016/j.jneuroling.2009.05.002.

Elvevåg, B., Foltz, P.W., Rosenstein, M., Ferrer-i-Cancho, R., De Deyne, S., Mizraji, E., Cohen, A., 2017. Thoughts about disordered thinking: measuring and quantifying the laws of order and disorder. Schizophr. Bull. 43 (3), 509–513. https://doi.org/10.1093/schbul/sbx040.

Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr. Res. 93 (1–3), 304–316. https://doi.org/10.1016/j.schres.2007.03.001.

Foltz, P.W., Chandler, C., Diaz-Asper, C., Cohen, A.S., Rodriguez, Z., Holmlund, T.B., Elvevåg, B., 2023. Reflections on the nature of measurement in language-based automated assessments of patients' mental state and cognitive function. Schizophr. Res. https://doi.org/10.1016/j.schres.2022.07.011.

Grabb, D., 2023. The impact of prompt engineering in large language model performance: a psychiatric example. J. Med. Artif. Intell. 6 https://doi.org/10.21037/jmai-23-71.

Guo, W., Caliskan, A., 2021. Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122–133. https://doi.org/10.1145/3461702.3462536.

Gupta, T., Hespos, S.J., Horton, W.S., Mittal, V.A., 2018. Automated analysis of written narratives reveals abnormalities in referential cohesion in youth at ultra high risk for psychosis. Schizophr. Res. 192, 82–88. https://doi.org/10.1016/j.schres.2017.04.025.

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., Cambria, E., 2023. A Survey of Large Language Models for Healthcare: From Data, Technology, and Applications to Accountability and Ethics. arXiv arXiv:2310.05694.

Hitczenko, K., Cowan, H., Mittal, V., Goldrick, M., 2021. Automated coherence measures fail to index thought disorder in individuals at risk for psychosis. In: Goharian, N., Resnik, P., Yates, A., Ireland, M., Niederhoffer, K., Resnik, R. (Eds.), Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access. Association for Computational Linguistics, pp. 129–150. https://doi.org/10.18653/v1/2021.clpsych-1.16.

Holmlund, T.B., Chandler, C., Foltz, P.W., Cohen, A.S., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Elvevåg, B., 2020. Applying speech technologies to assess verbal memory in patients with serious mental illness. NPJ Digit. Med. 3, 1–8. https://doi.org/10.1038/s41746-020-0241-7.

Iter, D., Yoon, J., Jurafsky, D., 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In: Loveys, K., Niederhoffer, K., Prud'hommeaux, E., Resnik, R., Resnik, P. (Eds.), Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic. Association for Computational Linguistics, pp. 136–146. https://doi.org/10.18653/v1/W18-0615.

Jastak, S.R., Wilkinson, G.S., 1984. *WRAT-R: Wide range Achievement Test-Revised Administration Manual* (1984 rev. Ed). Jastak Associates.

Jin, H., Chen, S., Wu, M., Zhu, K.Q., 2023. PsyEval: A Comprehensive Large Language Model Evaluation Benchmark for Mental Health. arXiv arXiv:2311.09189.

Just, S.A., Haegert, E., Koranova, N., Broecker, A.-L., Nenchev, I., Funcke, J., Heinz, A., Bermpohl, F., Stede, M., Montag, C., 2020. Modeling incoherent discourse in nonaffective psychosis. Front. Psychiatr. 11, 1–11. https://doi.org/10.3389/fpsyt.2020.00846.

Kircher, T., Krug, A., Stratmann, M., Ghazi, S., Schales, C., Frauenheim, M., Turner, L., Fährmann, P., Hornig, T., Katzev, M., Grosvald, M., Müller-Isberner, R., Nagels, A., 2014. A rating scale for the assessment of objective and subjective formal thought

and language disorder (TALD). Schizophr. Res. 160 (1), 216–221. https://doi.org/10.1016/j.schres.2014.10.024.
Kojima, T., Gu, S.(Shane, Reid, M., Matsuo, Y., Iwasawa, Y., 2022. Large language models are zero-shot reasoners. Adv. Neural Inf. Process. Syst. 35, 22199–22213.
Lamichhane, B., 2023. Evaluation of ChatGPT For NLP-based Mental Health Applications. arXiv. arXiv:2303.15727. http://arxiv.org/abs/2303.15727.
Li, H., Moon, J.T., Purkayastha, S., Celi, L.A., Trivedi, H., Gichoya, J.W., 2023. Ethics of large language models in medicine and medical research. Lancet Digit. Health 5 (6), e333–e335. https://doi.org/10.1016/S2589-7500(23)00083-3.
Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G., 2023. Pre-train, Prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM. Comput. Surv. 55 (9), 1–35. https://doi.org/10.1145/3560815.
Low, D.M., Bentley, K.H., Ghosh, S.S., 2020. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig. Otolaryngol. 5 (1), 96–116. https://doi.org/10.1002/lio2.354.
Luykx, J.J., Gerritse, F., Habets, P.C., Vinkers, C.H., 2023. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. World Psychiatr. 22 (3), 479–480. https://doi.org/10.1002/wps.21145.
Minssen, T., Gerke, S., Aboy, M., Price, N., Cohen, G., 2020. Regulatory responses to medical machine learning. J. Law Biosci. 7 (1) https://doi.org/10.1093/jlb/lsaa002.
Morgan, S.E., Diederen, K., Vértes, P.E., Ip, S.H.Y., Wang, B., Thompson, B., Demjaha, A., De Micheli, A., Oliver, D., Liakata, M., Fusar-Poli, P., Spencer, T.J., McGuire, P., 2021. Natural language processing markers in first episode psychosis and people at clinical high-risk. Transl. Psychiatr. 11 (1), 1. https://doi.org/10.1038/s41398-021-01722-y.
Mota, N.B., Vasconcelos, N.A.P., Lemos, N., Pieretti, A.C., Kinouchi, O., Cecchi, G.A., Copelli, M., Ribeiro, S., 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. PLoS. One 7 (4). https://doi.org/10.1371/journal.pone.0034928.
Naismith, B., Mulcaire, P., et al., 2023. Automated evaluation of written discourse coherence using GPT-4. In: Burstein, J., Kochmar, E., Burstein, J., Horbach, A., Laarmann-Quante, R., Madnani, N., et al. (Eds.), Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Eds. Association for Computational Linguistics, pp. 394–403. https://doi.org/10.18653/v1/2023.bea-1.32.
Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S.M., Ness, R.O., Poon, H., Qin, T., Usuyama, N., White, C., Horvitz, E., 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. arXiv. arXiv:2311.16452. http://arxiv.org/abs/2311.16452.
OpenAI Platform. (2023). Retrieved December 5, 2023, from https://platform.openai.com/docs/api-reference/chat.
Parola, A., Lin, J.M., Simonsen, A., Bliksted, V., Zhou, Y., Wang, H., Inoue, L., Koelkebeck, K., Fusaroli, R., 2023. Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of NLP automated measures of coherence. Schizophr. Res. 259, 59–70. https://doi.org/10.1016/j.schres.2022.07.002.
Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Bavel, J.J.V. (2023). *GPT is an effective tool for multilingual psychological text analysis*. https://doi.org/10.31234/osf.io/sekf5.
Rosenstein, M., Foltz, P.W., DeLisi, L.E., Elvevåg, B., 2015. Language as a biomarker in those at high-risk for psychosis. Schizophr. Res. 165 (2–3), 249–250. https://doi.org/10.1016/j.schres.2015.04.023.
Sadeghi, M., Egger, B., Agahi, R., Richer, R., Capito, K., Rupp, L.H., Schindler-Gmelch, L., Berking, M., Eskofier, B.M., 2023. Exploring the capabilities of a language model-only approach for depression detection in text data. In: 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1–5. https://doi.org/10.1109/BHI58575.2023.10313367.
Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., Okruszek, L., 2021. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Res. 304, 114135 https://doi.org/10.1016/j.psychres.2021.114135.
Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., Arcas, B.A.y, Webster, D., Natarajan, V., 2022. Large Language Models Encode Clinical Knowledge. arXiv. arXiv:2212.13138. http://arxiv.org/abs/2212.13138.
Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Arcas, B.A.y, Natarajan, V., 2023. Towards Expert-Level Medical Question Answering With Large Language Models. arXiv. arXiv:2305.09617. http://arxiv.org/abs/2305.09617.
Tang, S.X., Kriz, R., Cho, S., Park, S.J., Harowitz, J., Gur, R.E., Bhati, M.T., Wolf, D.H., Sedoc, J., Liberman, M.Y., 2021. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. NPJ. Schizophr. 7, 1–8. https://doi.org/10.1038/s41537-021-00154-3.
Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W., 2023. Large language models in medicine. Nat. Med. 29 (8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is All you Need. In: Advances in Neural Information Processing Systems, 30. In: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
Voleti, R., Woolridge, S.M., Liss, J.M., Milanovic, M., Stegmann, G., Hahn, S., Harvey, P. D., Patterson, T.L., Bowie, C.R., Berisha, V., 2023. Language analytics for assessment of mental health status and functional competency. Schizophr. Bull. 49 https://doi.org/10.1093/schbul/sbac176.
Voppel, A., de Boer, J., Brederoo, S., Schnack, H., Sommer, I., 2021. Quantified language connectedness in schizophrenia-spectrum disorders. Psychiatr. Res. 304, 114130 https://doi.org/10.1016/j.psychres.2021.114130.
Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., Yue, C., Zhang, H., Liu, Y., Li, X., Ge, B., Zhu, D., Yuan, Y., Shen, D., Liu, T., Zhang, S., 2023. Prompt engineering for healthcare: methodologies and applications. arXiv arXiv.2304.14670.
Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M., 2020. Generalizing from a few examples: a survey on few-shot learning. ACM. Comput. Surv. 53 (3) https://doi.org/10.1145/3386252.
Wechsler, D., 1981. WAIS-R: Manual : Wechsler Adult Intelligence Scale–Revised. Harcourt Brace Jovanovich [for] Psychological Corp.
Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A.K., Wang, D., 2023. Mental-LLM: Leveraging Large Language Models For Mental Health Prediction via Online Text Data. arXiv. arXiv:2307.14385. http://arxiv.org/abs/2307.14385.
Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., Ananiadou, S., 2023. Towards Interpretable Mental Health Analysis With Large Language Models. arXiv. arXiv:2304.03347. http://arxiv.org/abs/2304.03347.