Aiding LLMs with clinical scoresheets does not improve neurobehavioral diagnostic classifications from text using basic prompting

Kaiying Lin, Abdur Rasool, Saimourya Surabhi, Cezmi Mutlu, Haopeng Zhang, Dennis Wall Paul, Peter Washington

Submitted to: Journal of Medical Internet Research
on: March 28, 2025

Table of Contents

Aiding LLMs with clinical scoresheets does not improve neurobehavioral diagnostic classifications from text using basic prompting

Kaiying Lin[1]; Abdur Rasool[2]; Saimourya Surabhi[3]; Cezmi Mutlu[3]; Haopeng Zhang[2]; Dennis Wall Paul[3]; Peter Washington[4]

[1] Institute of Linguistics Academia Sinica Taipei TW
[2] University of Hawaii System Honolulu US
[3] Stanford University Stanford US
[4] University of California, San Francisco San Francisco US

Corresponding Author:
Peter Washington

University of California, San Francisco
505 Parnassus Ave
San Francisco
US

Abstract

Background: Large language models (LLMs) have demonstrated the ability to perform complex tasks traditionally requiring human intelligence. However, their use in automated diagnostics for psychiatry and behavioral sciences remains understudied.

Objective: To evaluate whether simple prompting of LLM-based chatbots can support diagnostic classification for neuropsychiatric conditions, and whether providing clinical assessment scales improves performance.

Methods: We tested two approaches using ChatGPT and Claude: (1) direct diagnostic querying and (2) execution of chatbot-generated code for classification. Three diagnostic datasets were used: ASDBank (autism), AphasiaBank (aphasia), and DAIC-WOZ (depression and related conditions). Each approach was evaluated with and without the aid of clinical assessment scales. Performance was compared to existing machine learning benchmarks on these datasets.

Results: Across all three datasets, incorporating clinical assessment scales led to modest improvements in performance, but results remained inconsistent and below the performances in prior studies. On the AphasiaBank dataset, the Direct Diagnosis approach with ChatGPT 4 yielded a relatively low F1 score (0.655) and very low specificity (0.33). Using the Code Generation method, performance improved substantially, with ChatGPT 4o achieving an F1 score of 0.814, specificity of 0.786, and sensitivity of 0.843. On the ASDBank dataset, Direct Diagnosis approaches yielded lower F1 scores (0.598 for ChatGPT 4 and 0.514 for ChatGPT 4o), while the Code Generation approach with Claude 3.5 improved performance to an F1 score of 0.6, specificity of 0.67, and sensitivity of 0.69. In the Daic-woz dataset, the Direct Diagnosis method produced high sensitivity (0.939) but very low specificity (0.08) and an F1 score of 0.452 with ChatGPT 4. Code Generation improved specificity (up to 0.886 with ChatGPT 4o), but F1 scores remained low overall, ranging from 0.203 to 0.33. These findings indicate that while clinical scales can help structure outputs in both approaches they do not consistently enable LLMs to reach clinically high diagnostic performance when using simple prompts.

Conclusions: Current LLM-based chatbots, when prompted naïvely, underperform on psychiatric and behavioral diagnostic tasks compared to specialized machine learning models. Clinical assessment scales might modestly aid chatbot performance, but more sophisticated prompt engineering and domain integration are likely required to reach clinically actionable standards. Clinical Trial: Not applicable.

(JMIR Preprints 28/03/2025:75030)
DOI: https://doi.org/10.2196/preprints.75030

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
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Original Paper

Kaiying Kevin Lin[1,2,] Abdur Rasool[1], Saimourya Surabhi[3], Cezmi Mutlu[3], Haopeng Zhang[1], Dennis Paul Wall[3], Peter Y. Washington[4]*

[1] Department of Information and Computer Science, University of Hawai'i, Manoa, Honolulu, HI, USA

[2] Institute of Linguistics, Academia Sinica, Taipei, Taiwan

[3] Department of Pediatrics, Division of Clinical Informatics, Stanford University, Stanford, CA, USA

[4] Division of Clinical Informatics and Digital Transformation, Department of Medicine, University of California - San Francsico, San Francisco, CA, USA

Aiding LLMs with clinical scoresheets does not improve neurobehavioral diagnostic classifications from text using basic prompting

Abstract

Background: Large language models (LLMs) have demonstrated the ability to perform complex tasks traditionally requiring human intelligence. However, their use in automated diagnostics for psychiatry and behavioral sciences remains understudied.

Objective: To evaluate whether simple prompting of LLM-based chatbots can support diagnostic classification for neuropsychiatric conditions, and whether providing clinical assessment scales improves performance.

Methods: We tested two approaches using ChatGPT and Claude: (1) direct diagnostic querying and (2) execution of chatbot-generated code for classification. Three diagnostic datasets were used: ASDBank (autism), AphasiaBank (aphasia), and DAIC-WOZ (depression and related conditions). Each approach was evaluated with and without the aid of clinical assessment scales. Performance was compared to existing machine learning benchmarks on these datasets.

Results: Across all three datasets, incorporating clinical assessment scales led to modest improvements in performance, but results remained inconsistent and below the performances in prior studies. On the AphasiaBank dataset, the Direct Diagnosis approach with ChatGPT 4 yielded a relatively low F1 score (0.655) and very low specificity (0.33). Using the Code Generation method, performance improved substantially, with ChatGPT 4o achieving an F1 score of 0.814, specificity of 0.786, and sensitivity of 0.843. On the ASDBank dataset, Direct Diagnosis approaches yielded lower F1 scores (0.598 for ChatGPT 4 and 0.514 for ChatGPT 4o), while the Code Generation approach with Claude 3.5 improved performance to an F1 score of 0.6, specificity of 0.67, and sensitivity of 0.69. In the Daic-woz dataset, the Direct Diagnosis method produced high sensitivity (0.939) but very low specificity (0.08) and an F1 score of 0.452 with ChatGPT 4. Code Generation improved specificity (up to 0.886 with ChatGPT 4o), but F1 scores remained low overall, ranging from 0.203 to 0.33. These findings indicate that while clinical scales can help structure outputs in both approaches they do not consistently enable LLMs to reach clinically high diagnostic performance when using simple prompts.

Conclusions: Current LLM-based chatbots, when prompted naïvely, underperform on psychiatric and behavioral diagnostic tasks compared to specialized machine learning models. Clinical assessment scales might modestly aid chatbot performance, but more sophisticated prompt engineering and domain integration are likely required to reach clinically actionable standards.

Trial Registration: Not applicable.

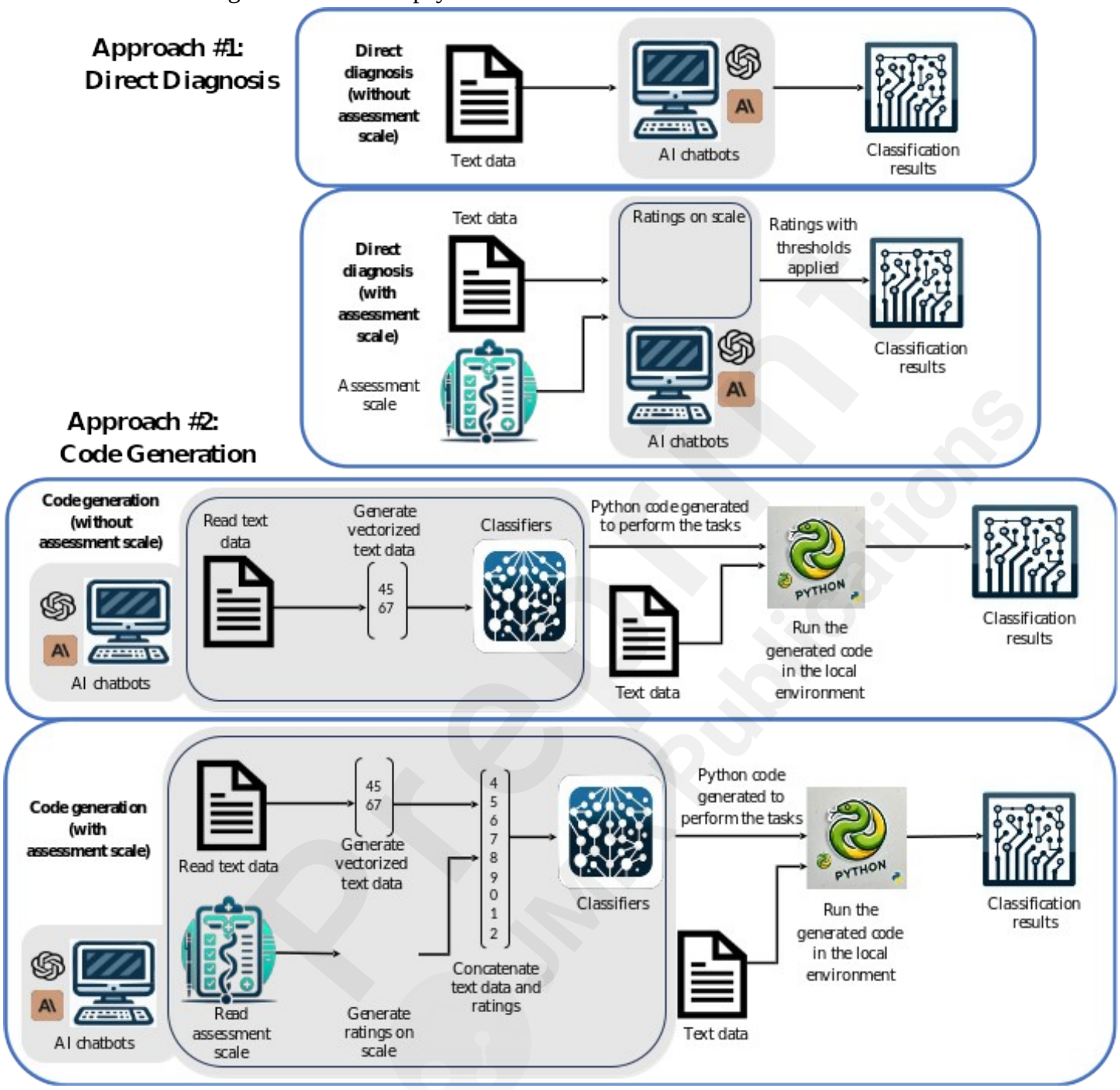
Keywords: neurological diagnostics, classification, LLM, chatbot

Introduction

Large language models (LLMs) have recently demonstrated capabilities that closely approximate or exceed human cognitive functions in various domains [1-3]. Given their efficacy in executing complex tasks, there is a burgeoning interest in exploring the potential applications of LLMs in clinical settings, including in areas such as providing emotional support [4-5] and mental health diagnoses [6-10]. The diagnostic process for neurobehavioral conditions typically encompasses comprehensive clinical assessments and longitudinal behavioral observations [11-12]. The integration of LLMs (as well as other ML models) into this process could potentially streamline this complex and time-consuming diagnostic procedure by facilitating automated screening processes [13-14]. While some research indicates that the proficiency of LLMs in these tasks remains modest [15-18], they are emerging as a promising avenue for developing scalable and accessible screening services.

ChatGPT [19] and Claude [20], prominent LLM-based conversational agents, have been the subject of evaluation for their potential in digital neurobehavioral diagnostics. Previous studies have indicated that the capabilities of chatbots for neurobehavioral classification remain limited, even when assessing specific conditions or smaller patient cohorts [10, 21-22]. In response to these findings, subsequent research efforts have focused on enhancing ChatGPT's performance in this domain [6, 21], typically employing varied prompting strategies such as the formulation of precise inquiries and the provision of relevant contextual information.

Building upon these foundational works, we aim to evaluate the use of LLM-based chatbots to aid in automated diagnostics for neuropsychiatric conditions with the aid of assessment scales. We evaluate



two paradigms: (1) directly deriving diagnoses from textual data and (2) chatbot-generated code executed in a local environment for diagnostic classification. As an attempt to reach clinical relevance, we instructed the chatbots to either provide ratings on standardized clinical assessment scales or incorporate these ratings into their diagnostic algorithms. This approach aims to leverage established clinical metrics while harnessing the analytical capabilities of LLMs. We hypothesized that offering this clinically grounded method for automated diagnosis would lead to improved diagnostic performances.

Methods

We implemented four distinct methodologies to evaluate the diagnostic capabilities of the chatbots

(Figure 1). In the direct diagnosis approach without assessment scales, we directly input data to the conversational AI model, which then generates classification results. This process involved providing the chatbot with the processed dataset as input data and defining its primary task as emitting neurobehavioral classification results for the condition of interest. We instructed the chatbot to derive diagnostic classifications for all participants using the text data from the entire processed dataset. If the chatbot indicated an inability to perform this task and requested to run in our environment, we directed it to utilize its pretrained knowledge and complete the task (see the Appendix for the detailed prompts of each condition).

The direct diagnosis approach with assessment scales involved inputting both data and a clinical assessment scale into the model; the model was subsequently tasked with rating items on the scale and provided these ratings as output. We then applied predefined thresholds from the clinical assessment scales to each subject's ratings to derive final neurobehavioral diagnoses.

For both direct diagnosis conditions, we performed zero-shot classification without conducting train and test splits, as we aimed to evaluate the models' ability to generalize from their pretrained knowledge. This process was repeated five times for each condition, with results averaged across iterations to improve the robustness of our results.

In addition to direct diagnosis, we explored a code generation approach to determine whether LLM chatbots could perform automated neurobehavioral classification by externalizing their reasoning into executable models. The motivation for the code generation condition stems from the observation that large language models (LLMs) like ChatGPT often struggle with directly solving complex reasoning tasks but can excel at generating code that reliably solves these problems. A notable example is Sudoku: while ChatGPT frequently fails to correctly solve Sudoku puzzles through direct reasoning, it can generate Python code that solves them efficiently when executed. Recent studies have examined this phenomenon, revealing that LLMs perform poorly on complex logic-based tasks like Sudoku when relying solely on their internal reasoning capabilities, yet demonstrate improved performance when prompted to generate code that encodes the required logic [23-24]. This suggests that the reasoning and structure embedded in generated code may enable LLMs to circumvent some of the limitations they face during direct response generation. While the diagnostic processes under the code generation condition differ fundamentally from direct diagnosis, they offer a complementary perspective on the model’s capabilities. Notably, chatbots frequently produced executable code and asserted its effectiveness—even without explicit prompting in early iterations—highlighting the importance of formally evaluating this behavior.

In the code generation approach without assessment scales, which served as a control condition, we fed the processed data as input into the conversational AI model. Following the data review, we instructed the chatbot to select what it deemed the most appropriate algorithm for the task and to output the corresponding Python code. This code was subsequently executed in an external Python environment. We tasked the chatbot with conducting stratified 5-fold cross-validation on the dataset, reporting F1 score, specificity, sensitivity, and accuracy as performance metrics. To optimize results, we engaged in an iterative process with the chatbot, requesting performance improvements until the generated code produced results consistent with its previous two iterations.

The code generation approach with assessment scales began with providing the chatbot with the processed data as input followed by a standardized assessment scale. We then prompted the model to analyze the text and to apply established cutoff thresholds from these scales as output to determine the final diagnosis. However, observing that this often resulted in unsatisfactory classifications, we next encouraged the chatbots to incorporate these ratings into a machine-learning algorithm of their

own design. The chatbot emitted the algorithm, which we then ran in our local environment. If no further performance improvements from the “without assessment-scales” condition was observed, we directed the chatbot to revert to the algorithm used in the condition without the assessment scale and to integrate the assessment scale ratings into the training procedures. This methodology facilitated a direct comparison of performance between two code generation approaches, making any potential improvements attributable to the assessment scale approach.

Table 1 displays the final algorithm used in each code generation condition. We ensured that the chatbot incorporated the assessment scale ratings in its algorithms, requesting integration if initially omitted. The iterative process for each condition continued until the performance of the generated code reached a plateau, with no significant improvement observed over two consecutive iterations.

Table 1. Machine learning approaches emitted by the chatbots in the two code generation conditions (LR: Logistic Regression, RF: Random Forest).

	Database		
	AphasiaBank	**ASDBank**	**Daic-woz Database**
Code Generation (No Assessment Scale) Approach			
Chatgpt4	Tf-idf+LR	Tf-idf+LR	CountVectorizer+LR
Chatgpt4o	Wordcount	CountVectorizer+LR	Tf-idf + XGBClassifier
Claude 3.5	Wordcount	Tf-idf+LR	Tf-idf +LR
Code Generation (Assessment Scale) Approach			
Chatgpt4	Tf-idf+LR	Tf-idf+LR	CountVectorizer+LR
Chatgpt4o	Wordcount+LR	CountVectorizer+LR	Tf-idf + XGBClassifier
Claude 3.5	Wordcount+threshold	Tf-idf+RF	Tf-idf +LR

Datasets

We used three distinct databases, each focusing on a specific neurobehavioral condition. Two of these, the ASDBank [27] and AphasiaBank [28], are sourced from TalkBank (https://www.talkbank.org) and contain language samples for autism and aphasia, respectively, while the third, the Daiz-woz database [29], contains textual data from patients with depression, anxiety, and PTSD.

AphasiaBank [28] is a repository containing multimedia language samples from both aphasia and control participants. These samples were collected through standardized discourse tasks, including unstructured speech samples, picture descriptions, story narratives, and procedural discourse.

ASDBank [27] comprises a collection of language samples and interactions from individuals diagnosed with autism. The data within ASDBank includes transcribed audio and video recordings of clinical interviews and naturalistic interactions.

We utilized all available English transcripts from both AphasiaBank and ASDBank. Data processing was performed to consolidate all samples from a single participant into one datapoint. The resulting dataset comprises 715 aphasia datapoints and 352 control datapoints for AphasiaBank, and 34 ASD and 44 control datapoints for ASDBank.

The Daic-woz database [28] consists of semi-structured interviews conducted by a virtual agent designed to identify symptoms of depression and PTSD. These interviews include questions about personal experiences, quality of life, and emotions. We consolidated all samples from a participant, including the interviewer's input, into a single datapoint. The Daic-woz database includes 56 patient datapoints and 133 control datapoints.

Table 2 provides a summary of the diagnosis distribution across each dataset.

Table 2. The numbers of control and patient datapoints in each of the datasets we evaluated.

DataBank	**Control**	**Patient**

AphasiaBank	352	715
ASDBank	44	34
Daic-woz	133	56

Aphasia, depression, and autism each manifest distinct linguistic characteristics that are both overlapping and unique. Aphasia, typically resulting from brain damage, is characterized by impaired language production and comprehension, often including repetitive language and the frequent use of filler words as individuals struggle to retrieve or organize words effectively [25]. Depression, while primarily a mood disorder, affects language through reduced verbal output, monotone speech, and a preference for negative or self-critical language patterns. Depressive language, such as expressions of negativity, can be a key symptom of the condition. Another characteristic linguistic feature is an excessive number of sighs, reflecting physical or emotional fatigue. Autism is marked by unique communication challenges, including delayed speech development, echolalia (repetition of phrases), difficulty with pragmatic language (e.g., understanding sarcasm or social cues), and overly literal or formal speech. Autistic individuals may also exhibit fragmented sentences and frequent use of filler words, reflecting challenges in organizing thoughts or navigating social interactions [26].

Many previous studies have leveraged the datasets we used in our research. However, much of the existing work has focused on advanced tasks such as multimodal detection or severity classification rather than simpler text-based binary classification using chatbots. These studies have often achieved strong (though not clinically translatable) performances, frequently exceeding 80% in F1 scores or accuracy. For example, Dinkel et al. applied a text-based multi-task network to the Daic-woz dataset, achieving an F1 score of 0.86 for severity detection [30]. Similarly, Agrawal et al. employed a fused BERT-BiLSTM model integrated with XGBoost to perform binary classification, achieving an F1 score of 91% [31].

For the AphasiaBank dataset, most prior studies have focused on severity classification, making direct comparisons with our binary classification studies challenging. The only relevant work by Cong et al. found that using LLM-derived surprisal features facilitated detection, achieving 79% in both accuracy and F1 score [30]. Similarly, studies involving the ASDBank dataset are limited, partly due to the recent development of the dataset. Chu et al. included another dataset, CHILDES, as a source of healthy control data. By extracting a few linguistic features from these two datasets, their binary classification approach reached an F1 score of 80% [33].

These studies suggest that LLM-based models directly diagnosing from the datasets used in this study should achieve high performance if chatbots exhibit comparable classification capabilities to fine-tuned LLMs.

Models

We evaluated two approaches using three state-of-the-art conversational AI models: GPT-4, GPT-4o and Claude 3.5 Sonnet. These models were selected because they are some of the most widely-used modern LLMs and because their efficacy in neurobehavioral classification tasks remain underexamined in current literature.

2.3. Assessment scales

We incorporated three widely recognized assessment scales and checklists used in clinical settings. We selected scales that assess behaviors at least tangentially related to language and that do not require extended observation periods. For example, the Autism-Spectrum Quotient (AQ) evaluates traits such as social preferences (“S/he prefers to do things with others rather than on her/his own”), behavioral patterns (“S/he prefers to do things the same way over and over again”), and attention capabilities (“have difficulty sustaining attention in tasks or fun activities”). The rating system for

this checklist—'definitely disagree', 'slightly disagree', 'slightly agree', and 'definitely agree'—does not necessitate longitudinal observation, unlike scales that use time-sensitive ratings such as 'rarely', 'less often', 'very often', and 'always'.
The assessment scales and checklists included in our study are as follows:

AphasiaBank: the Fluency test in WAB-AQ Western Aphasia Battery-Aphasia Quotient [34]
ASDBank: Autism-Spectrum Quotient (AQ) [35]
Daic-woz database: Burn's Depression Checklist [36]

In the two direct diagnosis conditions, we conducted the experimental procedure 5 times and obtained results based on the entirety of each dataset. We did not perform a train-and-test split for these conditions, instead opting for a zero-shot classification approach to assess the models' ability to generalize from their trained knowledge. In the code generation conditions, however, we instructed the chatbot to perform stratified 5-fold cross-validation on the entire dataset. The train and test split ratio during each fold was 4:1. Results were evaluated based on the test sets generated during each fold and subsequently averaged. Thus, the results from the direct diagnosis and code generation conditions are not directly comparable with each other.

Results

Core Results

Tables 3 – 8 and Figures 2 - 4 show the cross-validation results of two approaches for each dataset. We reported accuracy, F1 scores, specificity, and sensitivity. The results from the direct diagnosis conditions show varied performance across datasets.

Tables 3-4 and Figure 2 compare approaches on the AphasiaBank dataset, compared against Cong et al.'s baseline performance of 79% across metrics [30]. By contrast, our Direct Diagnosis condition yielded lower performance, particularly for ChatGPT 4 (F1: 0.655, specificity: 0.33). Our Code Generation condition improved results significantly, with ChatGPT 4o achieving the highest F1 score (0.814) and balanced specificity (0.786) and sensitivity (0.843), surpassing Cong et al.'s baseline.

The results on the ASDBank dataset are compared against Chu et al.'s baseline [33] which achieved an F1 score of 0.85 and a high sensitivity of 0.94, though specificity was notably low at 0.2. Our Direct Diagnosis approaches struggled in comparison, with ChatGPT 4 and 4o producing lower F1 scores (0.598 and 0.514, respectively) and poor specificity. The Code Generation condition improved overall performance, with Claude 3.5 achieving the highest accuracy (0.68) and balanced F1 (0.6), specificity (0.67), and sensitivity (0.69). ChatGPT 4 and 4o also showed improvement, but their performance was less consistent.

For the Daic-woz dataset, Dinkel et al. and Agrawal et al. set high baselines [30-31], achieving F1 scores of 0.84 and 0.91, respectively, with strong accuracy and sensitivity. In comparison, Direct Diagnosis approaches performed inconsistently, with ChatGPT 4 achieving a high sensitivity (0.939) but poor specificity (0.08) and F1 score (0.452), while ChatGPT 4o showed slightly better balance but still fell short. Code Generation approaches improved specificity, with ChatGPT 4o reaching 0.886 and Claude 3.5 achieving 0.767, but F1 scores remained low (0.203–0.33).

Overall, our findings reveal a substantial gap between two different approaches: Code Generation and Direct Diagnosis. While code generation generally led to improved performance compared to direct prompting, it still did not reach the levels reported in previous studies. Both approaches fell

short of established benchmarks, underscoring the limitations of current LLM-based diagnostic methods that rely solely on prompting without model fine-tuning.

Table 3. Results of four approaches on AphasiaBank using the Direct Diagnosis Approach.

AphasiaBank				
	Accuracy	**F1 score**	**Specificity**	**Sensitivity**
Cong et al. [32]	**0.79**	**0.79**		**0.79**
		Direct Diagnosis		
No Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.567±0.1	0.6556±0.136	0.33±0.3	0.684±0.29
ChatGPT 4o	0.561±0.029	0.648±0.111	0.397±0.11	0.642±0.22
Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.293±0.34	0.358±0.376	0.297±0.187	0.647±0.09
ChatGPT 4o	0.497±0.01	0.55±0.02	0.577±0.02	0.458±0.02

Table 4. Results of four approaches on AphasiaBank using the Code Generation Approach.

AphasiaBank				
		Code Generation		
No Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.67±0.16	0.74±0.17	0.79±0.24	0.40±0.31
ChatGPT 4o	0.67±0.0113	0.802±0.008	0.68±0.011	1
Claude 3.5	0.605 ± 0.034	0.623±0.036	0.844±0.037	0.488±0.033
Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.67±0.16	0.74±0.17	0.80±0.25	0.41±0.30
ChatGPT 4o	0.741±0.022	0.814±0.016	0.786±0.024	0.843±0.007
Claude 3.5	0.608± 0.036	0.627 ± 0.039	0.844 ± 0.037	0.492 ± 0.036

Figure 2. Accuracy and F1-score across conditions on the AphasiaBank dataset.

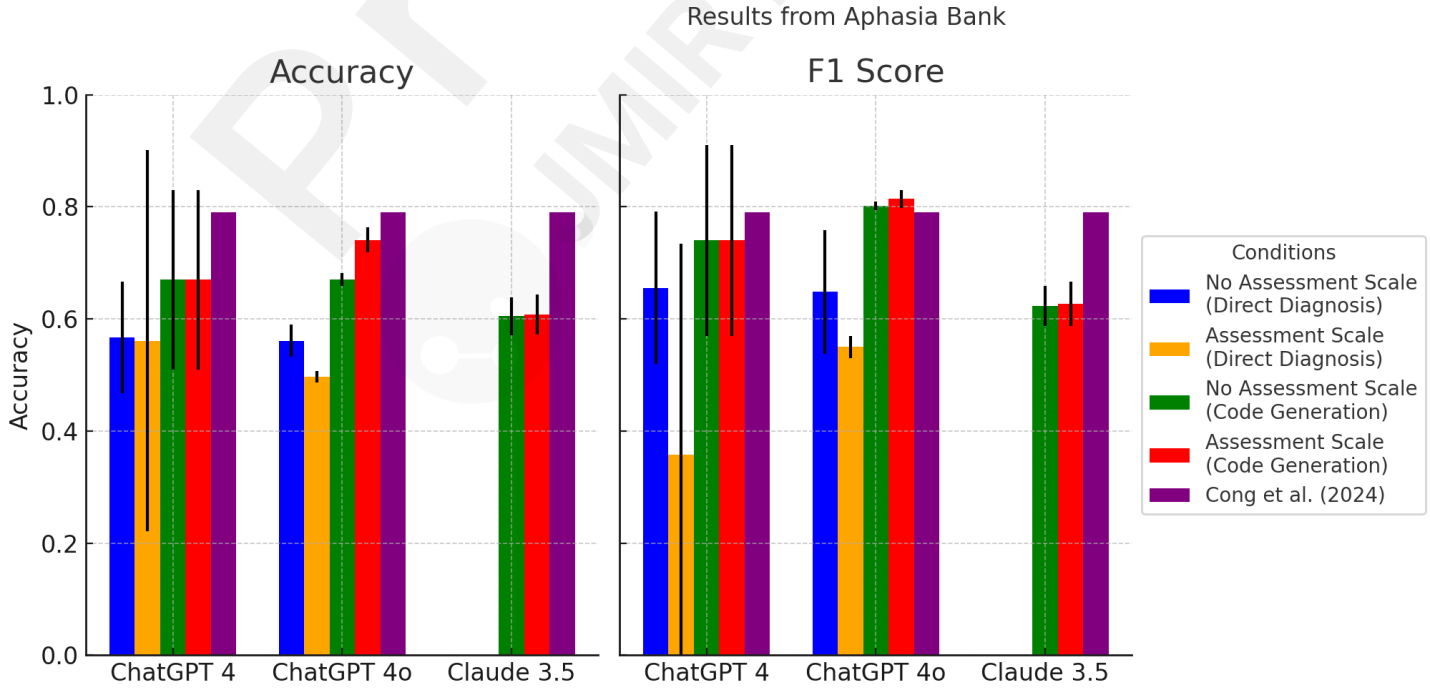


Table 5. Results of four approaches on ASDBank using the Direct Diagnosis Approach.

ASDBank				
	Accuracy	**F1 score**	**Specificity**	**Sensitivity**
Chu et al. [33]	**0.76**	**0.85**	**0.2**	**0.94**
Direct Diagnosis				
No Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.5±0.00	0.598±0.00	0.227±0.00	0.853±0.00
ChatGPT 4o	0.421±0.03	0.514±0.129	0.155±0.212	0.765±0.323
Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.427± 0.01	0.56±0.08	0.09±0.157	0.863±0.24
ChatGPT 4o	0.491±0.09	0.542±0.117	0.236±0.39	0.802±0.342

Table 6. Results of four approaches on ASDBank using the Code Generation Approach.

ASDBank				
Code Generation				
No Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.618±0.125	0.616±0.104	0.55±0.286	0.71±0.199
ChatGPT 4o	0.653±0.103	0.55±0.184	0.73±0.303	0.576±0.378
Claude 3.5	0.68 ±0.16	0.6±0.22	0.67±0.35	0.69±0.4
Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.642± 0.165	0.628±0.17	0.6±0.334	0.695±0.231
ChatGPT 4o	0.628±0.194	0.592±0.1974	0.689±0.325	0.578±0.257
Claude 3.5	0.64± 0.13	0.6 ± 0.23	0.69 ± 0.41	0.67 ± 0.36

Figure 3. Accuracy and F1-score across conditions on ASDBank.

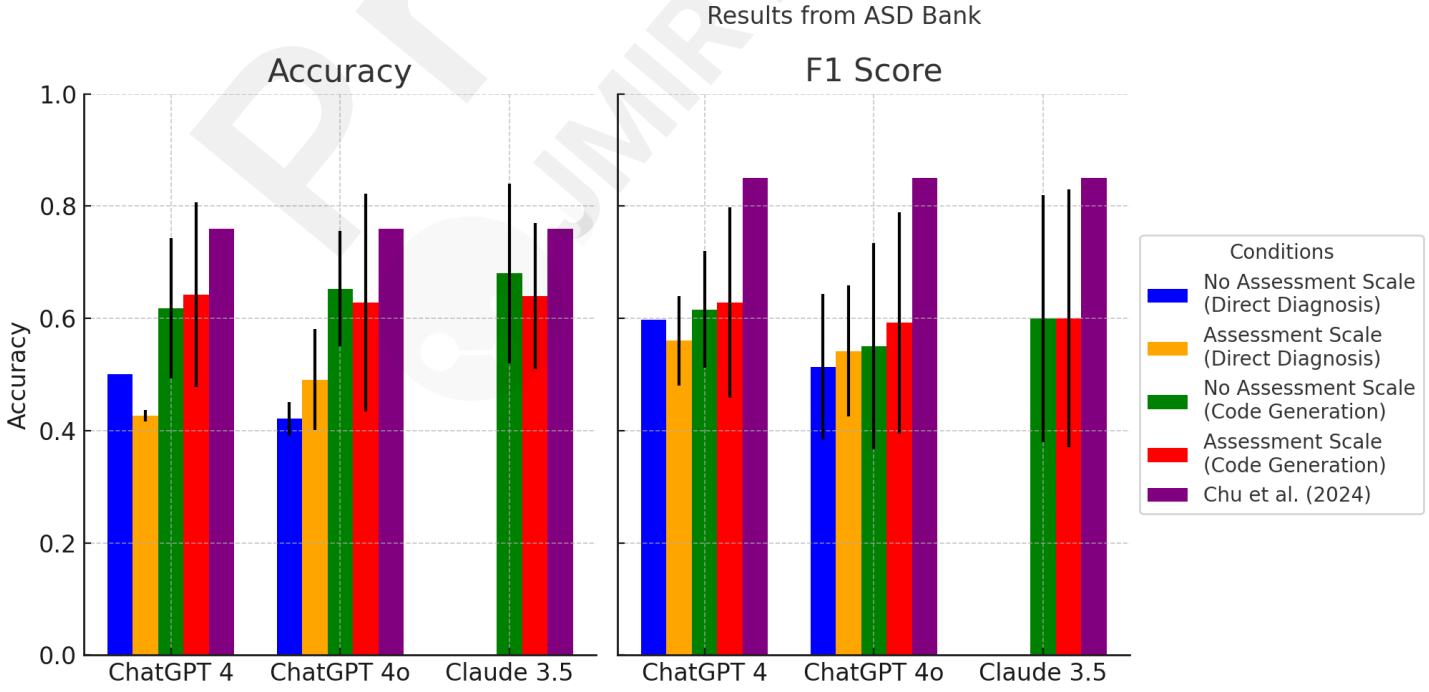


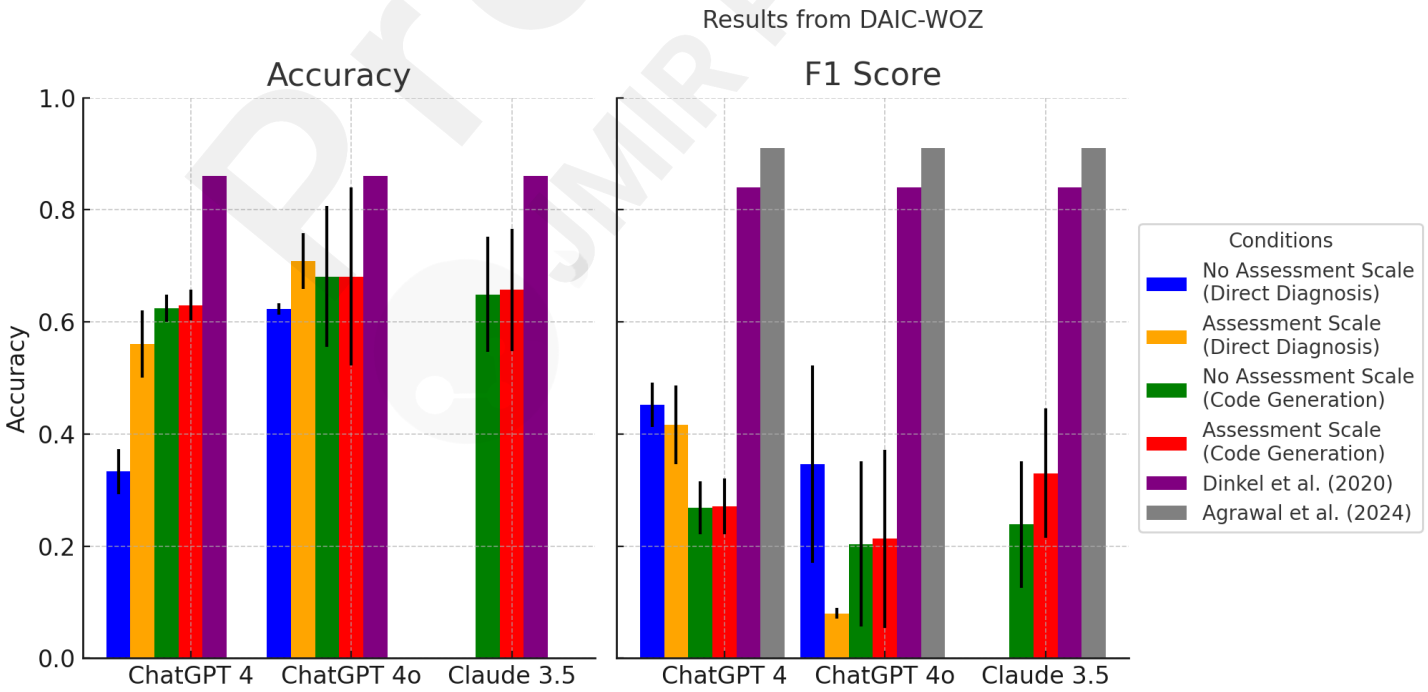
Table 7. Results of four approaches on the Daic-woz database using the Direct Diagnosis Approach.

Daic-woz Database				
	Accuracy	**F1 score**	**Specificity**	**Sensitivity**
Dinkel et al. [30]	**0.86**	**0.84**		**0.83**
Agrawal et al. [31]		**0.91**		**0.89**
Direct Diagnosis				
No Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.333± 0.04	0.452±0.04	0.08±0.51	0.939±0.039
ChatGPT 4o	0.623±0.01	0.346±0.176	0.711±0.168	0.409±0.347
Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.56± 0.06	0.416±0.07	0.56±0.08	0.516±0.05
ChatGPT 4o	0.709±0.05	0.08±0.01	1±0.00	0.429±0.006

Table 8. Results of four approaches on the Daic-woz database using the Code Generation Approach.

Daic-woz Database				
Code Generation				
No Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.624± 0.024	0.268±0.047	0.79±0.035	0.233±0.048
ChatGPT 4o	0.681±0.126	0.2038±0.1474	0.886±0.087	0.2286±0.2382
Claude 3.5	0.649± 0.103	0.2386±0.1131	0.7672±0.0251	0.2136±0.1131
Assessment Scale	**Accuracy**	**F1 score**	**Specificity**	**Sensitivity**
ChatGPT 4	0.63± 0.027	0.271±0.05	0.797±0.036	0.233±0.048
ChatGPT 4o	0.681±0.1587	0.213±0.1587	0.9±0.073	0.223±0.2389
Claude 3.5	0.657± 0.109	0.33 ± 0.1153	0.7738 ± 0.1	0.328 ± 0.1153

Figure 4. Accuracy and F1-score across conditions on the Daic-woz database.



Error Analysis

We first address the errors in the direct diagnosis approach, which did not appear to work well. We

observed that most rounds of classification yielded close-to-random performances for this classification. Interestingly, we noticed patterns in the emitted classification ratings, such as digits limited to only multiples of 3 or repeating sequences (e.g., 3, 2, 1, 0, 3, 2, 1, 0). We present the percentage of rounds, over 5 rounds of classification, that followed such patterns in Table 6. This demonstrates that a direct diagnosis prompting strategy does not work well for ChatGPT 4 and 4o.

Table 6. Percentage of random predictions.

Database	Approach	ChatGPT 4	ChatGPT 4o
AphasiaBank	wo-assessment-scale	80%	60%
	w-assessment-scale	20%	60%
ASDBank	wo-assessment-scale	20%	100%
	w-assessment-scale	100%	100%
Daic-woz Database	wo-assessment-scale	40%	80%
	w-assessment-scale	20%	60%

For the code generation approach, we found some examples of text archetypes (i.e., typical examples) that were frequently misclassified. These archetypes often reflect characteristics of the conditions. Common errors we observed included:

1. **Repetitive Language and Filler Words, Aphasia:** The presence of repetitive language patterns and an increased frequency of filler words led to misclassification in 100% of false positives for aphasia. Control participants' responses typically exhibited minimal repetition and filler word usage. However, even a slight elevation in these linguistic elements frequently resulted in misclassification, with the chatbots erroneously classifying control participants as positives. Notably, all misclassified false positives from all the chatbots contained these features.
2. **Fragmented Sentences and Filler Words, Autism:** Transcripts containing filler words or fragmented sentences were misclassified in almost 100% of cases as false positives originating from autistic individuals. With ChatGPT 4 and ChatGPT 4o, this archetype was observed in 100% of false positive data points, indicating a consistent misclassification pattern. In contrast, Claude 3.5 exhibited a different trend, because the majority of misclassified points were false negatives. Claude 3.5 did not appear to excessively utilize the linguistic features characteristic of this archetype.
3. **Lack of Depressive Language, Depression:** Text lacking overt depressive indicators and conveying generally positive sentiments accounted for about 80% of false negatives. For instance, statements such as "uh I'd say maybe the fact that it's a lot different than it was about ten years ago" and "I am pretty happy with the level of education I've gotten" often led to false negatives.
4. **Excessive Number of Laughter, Depression:** Text containing instances of laughter was classified as false negatives in about 90% of cases, originating from control subjects rather than individuals with depression.
5. **Excessive Number of Sighs, Depression:** Text containing references to sighing was categorized as false positives originating from individuals with depression. Over 50% of false positive cases included this feature, indicating its disproportionate influence on the classification process.

Table 7. Percentage of each text archetype in false positive/false negative datapoints in the code generation conditions, averaged across folds.

Archetype	Approaches	ChatGPT 4	ChatGPT 4o	Claude 3.5

1. Repetitive Language and Filler Words (false positives)	wo-assessment-scale	100%	100%	100%
	w-assessment-scale	100%	100%	100%
2. Fragmented Sentences and Filler Words (false positives)	wo-assessment-scale	100%	100%	0%
	w-assessment-scale	100%	100%	0%
3. Lack of Depressive Language (false negatives)	wo-assessment-scale	87.67%	89.02%	85%
	w-assessment-scale	87.67%	89.09%	79.09%
4. Excessive number of Laughter (false negatives)	wo-assessment-scale	88.36%	88.34%	90.7%
	w-assessment-scale	88.36%	88.9%	90.7%
5. Excessive Number of Sighs (false positives)	wo-assessment-scale	67.44%	69.23%	52%
	w-assessment-scale	65.12%	74.19%	52.17%

Table 7 details the frequency of occurrence of these archetypes. The observed misclassifications highlight the inherent constraints of relying on text-based methods for neurobehavioral diagnosis.

Discussion

This study reveals the limitations of using LLMs for automated neurobehavioral classification using text transcripts. Therefore, while human-in-the-loop processes are promising for such classification tasks [37-38], the human must likely remain more involved in the computational diagnosis procedure than simply prompting the LLM to generate a direct diagnosis, clinical rating, or classification code. In prior work for autism diagnostics, for example, humans extracted the behavioral features – a task that requires the ability to interpret relatively subjective human behavior – leaving the machine learning models to perform the simpler task of performing the final classification given the human-derived features [39-40]. It is likely that humans performing at least some level of analysis of the data will need to continue to be involved, and future prompt engineering approaches should explore these ideas more thoroughly.

In both direct-diagnosis conditions, we encountered significant limitations with both ChatGPT and Claude, which tended to generate random or close-to-random predictions. ChatGPT demonstrated an inability to provide accurate diagnoses. It occasionally refused to offer diagnoses, and when compelled to complete the tasks, the resulting classifications were not accurate. These challenges were even more pronounced with Claude 3.5, where we faced difficulties generating any classification results or ratings. The inclusion of assessment scales did not substantially improve performance, as the ratings on scale items also appeared to be randomly assigned. Notably, in many of these conditions, we observed a concerning trend where assessment scale ratings were often identical across participants, regardless of individual differences in their text data. Our attempts to prompt the models to improve their classifications or ratings by encouraging more careful text analysis did not yield significant improvements, and the results remained randomly generated.

Regarding the code generation condition, our findings indicate that LLM-generated machine learning pipelines show promising signs of improving diagnostic performance. However, these gains remain limited by the simplicity of the models typically produced. Chatbots often rely on basic methods when generating code, such as logistic regression and straightforward feature extraction, rather than employing more advanced or condition-specific techniques. Moreover, the selection of learning approaches by the chatbot tends to vary across iterations, often without clear rationale, suggesting a lack of consistency and optimization in the generated approaches. While the ability to generate functional code is encouraging, current LLMs still fall short in autonomously designing clinically appropriate and high-performing machine learning models for neurobehavioral diagnostics. These

limitations highlight the need for human oversight or more structured prompt strategies to better guide model generation toward clinically meaningful outcomes

It is important to note that previous studies have successfully achieved F1 scores of 70%-80% using subsets of the ASDBank, and high performance (F1 scores of 80-90%) using various methods on at least portions of the other two datasets (e.g., [30-33] [41-43]). By contrast, our results indicate that neither the direct diagnosis approaches nor the code generated by ChatGPT and Claude were able to attain similar results to these previous studies. This discrepancy suggests a gap between the optimal performance that chatbots can theoretically attain and the outcomes observed in our study. This may be due to our relatively straightforward methodological approach, which did not involve fine-tuning LLMs or employing more sophisticated state-of-the-art (yet general-purpose) prompting strategies.

In the conditions using assessment scales, we observed that the code from the chatbots did not apply diagnostic thresholds as defined by the assessment scales but instead directly incorporated the ratings as machine learning features. The rating methods were simplistic, frequently implemented a keyword-counting algorithm to provide ratings for ASDBank and Daic-woz. These ratings were then concatenated with features extracted from the feature extractor. The keyword-counting algorithm, which varied across conditions, typically involved identifying specific keywords relevant to the users' text data. For instance, in rating items on the ASD assessment scale, ChatGPT identified several keywords such as 'routine' and 'rude' from the assessment items to determine the ratings of the corresponding items on the assessment scale.

Limitations

We acknowledge several limitations of this study beyond the discussed performance gaps. (1) The scope of our investigation was constrained to three datasets, each representing distinct neurobehavioral conditions, with relatively few data points. This limits the robustness and generalizability of our findings and the ability of the models to learn. (2) Another limitation is the potential for a more judicious selection of clinical checklists and surveys for the assessment scale approaches guided by professional recommendations in the field of neurobehavioral diagnostics. This refinement in assessment tool selection could potentially yield improvements in the chatbots' performance. (3) Importantly, our prompting strategies were fairly limited. It is possible that more thorough state-of-the-art prompting strategies would have led to better performance. (4) It is likely that ChatGPT and Claude will be updated by the time this paper is published, reducing the reproducibility of this work. These limitations collectively underscore the preliminary nature of our findings and emphasize the need for more comprehensive investigation.

Recent advancements in LLM prompting techniques show great potential for enhancing the reasoning capability of LLMs, which may improve the neurobehavioral diagnostic accuracy in future studies building upon our study. For instance, chain-of-thought prompting, as demonstrated by Wei et al. (2023) [44], guides models to break down tasks into intermediate reasoning steps, achieving state-of-the-art performance on complex problems such as math word problems. This method has significant potential in neurobehavioral classification by allowing models to systematically analyze symptoms and diagnostic criteria step by step, improving diagnostic accuracy and reducing errors. Similarly, Madaan et al. (2023) [45] introduced SELF-REFINE, an iterative framework where LLMs evaluate and refine their outputs, resulting in a performance improvement of approximately 20% across various tasks. By refining previous prompts and generating explanations for their conclusions, this approach may enhance the transparency and reliability of diagnoses, fostering greater trust in AI-driven healthcare solutions. Furthermore, Wang et al. (2024) [46] proposed Solo Performance Prompting (SPP), which enables LLMs to collaborate with multiple simulated personas, thus reducing inaccuracies and enhancing problem-solving capabilities. In the context of healthcare, this

approach can mimic multidisciplinary diagnostic teams, offering diverse perspectives on complex cases and improving decision-making processes. Collectively, these methods highlight the potential of structured reasoning, iterative refinement, and cognitive synergy in advancing medical classifications, enhancing diagnostic accuracy, and broadening the utility of LLMs in medical and diagnostic applications. Future research may build upon these strategies to further improve LLM performance and expand their applicability in healthcare and beyond.

Conclusion

This study demonstrates that using naïve prompting strategies with popular LLM-based chatbots are inadequate for classifying neurobehavioral conditions from text transcripts, even when prompting the chatbot to leverage clinical assessment scales to focus its evaluation strategy. We recommend that future research explore more robust and complex prompting strategies, for example by integrating the assessment scale approach we explored here with state-of-the-art prompting strategies.

Acknowledgements

The authors gratefully acknowledge the grant support for the AphasiaBank project from the National Institutes of Health - National Institute on Deafness and Other Communication Disorders, grant number R01-DC008524 (2022-2027). We appreciate the permission to use ASDBank and Daic-woz Database. Additionally, the authors recognize the contributions of the ChatGPT and Claude language models in editing the grammar and phrasing of the manuscript and generating some of the icons used in the figures, though the authors have further edited the outputs of these models.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

The three databases used in this study can be accessed as follows:
AphasiaBank: https://aphasia.talkbank.org
ASDBank: https://asd.talkbank.org
Daic-woz Database : https://dcapswoz.ict.usc.edu

Multimedia Appendix

Direct Diagnosis (No Assessment Scale)

Me: Here's a data point from one of the participants in the dataset. Could you take a look at it?

Chatbot: It looks like you've shared a dialogue between a user, Ellie, and an automated conversational agent. This conversation covers a range of topics including personal in-terests, career goals, and some insights into Ellie's family dynamics and social habits. Is there something specific you're analyzing or looking for in this conversation?

Me: I now have additional data points from 189 users. Could you help classify these patients as having depression or not?

Chatbot: To classify the patients into those with depression and those without, we'll need to process

the text data you've provided, looking specifically for any explicit mentions of a depression diagnosis or symptoms strongly indicative of depression. From the snippets retrieved, here are examples of users being directly asked about depression and their responses:

1. Explicit confirmation of a depression diagnosis: "depression_diagnosed (have you been diagnosed with depression) yes".
2. Description of depression symptoms: "feeling hopeless all the time, not having enough energy to do things, feeling crummy about life".

Based on these and similar statements throughout the dataset, we can establish criteria for classification:

- Diagnosed with Depression: Any mention of being formally diagnosed by a pro-fessional.
- Symptoms Indicative of Depression: Descriptions of persistent sadness, loss of interest, fatigue, or other symptoms commonly associated with depression.

The next step would be to parse through each user's dialogue to flag these criteria. This can be automated with text analysis techniques to extract mentions of depression or its symptoms.

If you would like to proceed with an automated approach to classify these users based on the criteria, I can assist you in developing a script to process this text data. Let me know how you would like to proceed!

Me: Please choose an approach you regard as appropriate and provide a downloadable file.

Chatbot: I have processed the text data and classified the users based on mentions of depression or related symptoms. You can download the results from the following link:

(Continued with next iterations.)

Direct Diagnosis (Assessment Scale)

Me: Here's a data point from one of the participants in the dataset. Could you take a look at it?

Chatbot: Thanks for sharing this data! If you're looking to analyze or process this con-versation, could you let me know what specific aspects or tasks you're interested in ex-ploring? For instance, are you looking for sentiment analysis, thematic content extraction, or something else?

Me: Attached is Burn’s Depression Checklist along with the assessment scale to determine if a participant has depression:

(Include the assessment scale.)

Chatbot: Thanks for sharing the Burn's Depression Checklist and the scoring criteria. With this tool, you can assess individuals for various levels of depression based on their re-sponses to the 25-item questionnaire. Each item is rated on a scale from 0 to 4, indicating the severity or frequency of symptoms experienced over the past week.

If you're looking to apply this checklist to your dataset or need assistance in analyzing responses to determine the depression level of participants, let me know how I can assist further! Whether it's setting up a method to automate the scoring process or analyzing specific data entries, I'm here to

help.

Me: I now have a dataset with 189 participants. Could you provide the ratings for these 25 items for each participant and list them in an output file?

Chatbot: The dataset you've uploaded contains text responses from participants and does not explicitly list scores for the 25 items of Burn's Depression Checklist. If you want to extract and score responses for each item, you will need to manually or programmatically match the responses to the checklist items based on sentiment analysis or keyword matching.

Would you like assistance with developing an approach to score these responses based on the content of the text data?

Me: I would like you to apply your pre-trained knowledge to rate 25 items for each participant. Please compile all the ratings into a file that I can download.

Chatbot: To accurately score the Burn's Depression Checklist for each participant based on their responses, I would need to make inferences based on the emotional content and keywords within their responses. Since this involves a bit of interpretation, I will proceed with this approach:

1. Interpretation of Sentiments: I'll assess the sentiment in the responses related to each checklist item (e.g., feeling sad, discouraged, loss of interest) and assign scores ranging from 0 (Not At All) to 4 (Extremely).
2. Scoring: For each of the 25 items on the checklist, I'll infer a score based on the overall tone and content of the relevant sections of the conversation.
3. Output Preparation: The final scores for each participant will be compiled into a downloadable file.

Let me start by processing this data to generate the scores. This might take a moment.

I've generated the scores for the Burn's Depression Checklist for each participant and saved them in a file. You can download the file from the following link:
(The chatbot generates a file and continues to the next iteration.)

Code generation (No Assessment Scale)

Me: Here is a dataset file that I would like to use for a machine learning project. Could you please read the dataset? We will use this dataset soon.

Chatbot: I've successfully read the data from the four files you provided. Here's a brief overview of the first few rows of each dataset: (removed)… Is there anything specific you would like to do with this data?

Me: Alright, I want to perform detection of aphasia patients in this data. Can you perform a stratified 5-fold cross-validation on your method and also report F1 scores, sensitivity, specificity and standard deviations?

Chatbot: To perform a stratified 5-fold cross-validation and report metrics like F1 scores, sensitivity, specificity, and their standard deviations, we first need to define a clearer criterion or chatbot to classify participants as having potential signs of aphasia or not. Since we are working with textual data, we can create a simple classification system based on the metrics we've computed—like repetition rate and average sentence length—or use more advanced NLP features.

However, let's first create a simplistic threshold-based classifier, which will classify a participant as showing potential signs of aphasia if their repetition rate exceeds a certain threshold and their average sentence length is below another threshold.

(The chatbot introduces the code. The chatbot created the y_true label and did not use the actual labels).

Me: I have the true labels, what you call “y_true”, in a variable called “all_label” as well as the corresponding input text data in a variable called “all_data”, so you don't need to generate y_true labels by yourself. Could you please use my “all_label” variable?

Chatbot: Great! Since you have actual labels (all_label) and the participants' text data (all_data), we can use these directly for the classification and cross-validation process. The approach remains mostly the same, but we'll now use your real data and labels.

Me: (If the results are not good:) The results are very bad, can you please try another method?

(The chatbot introduces its algorithm and provides the code.)

Chatbot: If the initial simplistic method based on repetition rate and average sentence length didn't provide satisfactory results, it's likely that we need a more sophisticated approach to capture the nuances of aphasia in text data. For such cases, using Natural Language Processing (NLP) techniques can be much more effective.

Me: There seems to be a bug: TypeError: sequence item 0: expected str instance, list found.

Chatbot: The error you're encountering occurs because the data structure wasn't as I assumed. It appears that all_data might still be a list of lists of lists, where each partici-pant's data is a list of sentences, and each sentence is a list of words. We need to correctly flatten this to a list of strings for TF-IDF vectorization.

(The chatbot provides new code.)

(Continue refining the algorithm until it is free of bugs and delivers performance above random chance)

Code generation (Assessment Scale)

Me: Here is a dataset file that I would like to use for a machine learning project. Could you please read in the dataset? We will use this dataset soon.

Chatbot: I've successfully read the data from the four files you provided. Here's a brief overview of the first few rows of each dataset: (removed)… Is there anything specific you would like to do with this data?

Me: Alright, please provide the details of your methodology for conducting a stratified 5-fold cross-validation procedure to predict the presence of Autism Spectrum Disorder (ASD) based on the given dataset.

Chatbot: To perform Stratified 5-fold cross-validation on your data to predict if patients have ASD,

follow these steps using Python and scikit-learn:
1. Load and Preprocess Data: Ensure the data is loaded correctly and preprocessed to handle any missing values or categorical variables.
2. Define the Model: Set up a logistic regression model for binary classification.
3. Perform 5-Fold Cross-Validation: Use the cross_val_score function from scikit-learn to perform 5-fold cross-validation.

(The chatbot introduces the code)

Me: Would you please report F1 score, specificity and sensitivity, and their standard deviations?

Chatbot: To calculate F1 score, specificity, and sensitivity (recall) along with their standard deviations during Stratified 5-fold cross-validation, you can use the cross_val_score function for multiple metrics. Here's how you can modify the previous code to include these metrics:

(The chatbot introduces the code.)

Me: Now here's a list of the Autism-Spectrum Quotient (AQ) scale items. We aim to in-corporate these ratings into the classification model. According to the scale, users with a score exceeding 26 are considered to have ASD.

Chatbot: To incorporate the Autism Quotient (AQ) ratings into the code and calculate scores, we need to implement the scoring rules you provided. We'll integrate these scores into our data preprocessing step, and then use these scores as features in the model.
Here’s how to do it:
1. Define the scoring rules: Create a function to calculate AQ scores based on the re-sponses.
2. Include AQ scores in the dataset: Use these scores as additional features for the model.
3. Update the pipeline to include these features.

(The chatbot introduces the code.)

Ensure to replace the placeholder paths with the actual paths to your data files and adjust any specific details according to your dataset structure. This should now incorporate AQ scores into the model training and evaluation process.

Me: The text data provided does not include a pre-generated aq_responses variable. In-stead, we will need to programmatically derive the Autism-Spectrum Quotient (AQ) scale responses from the textual information available in the dataset.

Chatbot: Got it. We'll need to create a synthetic dataset for AQ responses to integrate with the main dataset. Here's the updated code to include the generation of AQ responses and their integration into the dataset:
1. Generate AQ responses: Create synthetic AQ responses for demonstration.
2. Calculate AQ scores: Use the provided AQ scoring rules.
3. Integrate AQ scores with text data: Use these scores as additional features for the model.

(The chatbot introduces the code.)

Me: Please refrain from generating synthetic ratings and instead focus on producing ratings derived directly from the provided text data.

Chatbot: I understand now. Here's a step-by-step approach to generate the responses based on the provided text data, calculate the AQ scores, identify participants with autism, and perform the stratified 5-fold cross-validation. Step-by-Step Implementation:

1. Extract Responses from text Data: Process the text data to extract responses.
2. Calculate AQ Scores: Calculate the AQ score for each participant based on their re-sponses.
3. Identify Participants with Autism: Determine which participants have an AQ score greater than 26.
4. Prepare Data for Cross-Validation: Use logistic regression to classify the data.
5. Perform Stratified 5-Fold Cross-Validation: Report the F1 score, specificity, sensitivity, and their standard deviations.

(The chatbot introduces the code.)

Me: The results indicate that the F1 scores and sensitivities are currently at zero. Could you please refine your methodology to enhance performance outcomes?

Chatbot: To avoid having a lot of zeros in the F1 scores and sensitivity, we can consider the following strategies:

1. Balance the Classes: Ensure that the dataset is balanced so that the model has enough samples of each class to learn from. This can be done through oversampling the minority class or undersampling the majority class.
2. Feature Engineering: Improve feature representation to make the data more in-formative for the model.
3. Model Tuning: Adjust hyperparameters of the model to better handle imbalanced data.

Here's an updated version of the code that includes class balancing using the SMOTE technique for oversampling the minority class:

(The chatbot introduces the code.)

(Continue refining the algorithm until it is free of bugs and delivers performance above random chance.)

References

1. Mathew, A. (2023). Is Artificial Intelligence a World Changer? A Case Study of OpenAI’s Chat GPT. Recent Progress in Science and Technology Vol. 5, 35–42. https://doi.org/10.9734/bpi/rpst/v5/18240D
2. Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. ArXiv, abs/2305.14627.
3. Biswas S. S. (2023). Potential Use of Chat GPT in Global Warming. Annals of biomedical engineering, 51(6), 1126–1127. https://doi.org/10.1007/s10439-023-03171-8
4. Lee, Y.K., Lee, I., Shin, M., Bae, S., & Hahn, S. (2023). Chain of Empathy: Enhancing Empathetic Response of Large Language Models Based on Psychotherapy Models. ArXiv, abs/2311.04915.
5. Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2023). Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2307.11991
6. Siyuan Chen, Zhiling Zhang, Mengyue Wu, and Kenny Zhu. (2023a). Detection of Multiple

Mental Disorders from Social Media with Two-Stream Psychiatric Experts. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9071–9084, Singapore. Association for Computational Linguistics

7. Chen, Z., Lu, Y., & Wang, W. Y. (2023b). Empowering Psychotherapy with Large Language Models: Cognitive Distortion Detection through Diagnosis of Thought Prompting. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2310.07146
8. Bhaumik, R., Srivastava, V., Jalali, A., Ghosh, S. & Chandrasekharan, R. (2023). Mindwatch: A smart cloud-based ai solution for suicide ideation detection leveraging large language models. medRxiv DOI: 10.1101/2023.09.25. 23296062 https://www.medrxiv.org/content/early/2023/09/26/2023.09.25.23296062.full.pdf.
9. Garg, R. K., Urs, V. L., Agarwal, A. A., Chaudhary, S. K., Paliwal, V., & Kar, S. K. (2023). Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. Health promotion perspectives, 13(3), 183–191. https://doi.org/10.34172/hpp.2023.22
10. Dergaa, I., Fekih-Romdhane, F., Hallit, S., Loch, A. A., Glenn, J. M., Fessi, M. S., Ben Aissa, M., Souissi, N., Guelmami, N., Swed, S., El Omri, A., Bragazzi, N. L., & Ben Saad, H. (2024). ChatGPT is not ready yet for use in providing mental health assessment and interventions. Frontiers in psychiatry, 14, 1277756. https://doi.org/10.3389/fpsyt.2023.1277756
11. Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three Approaches to Understanding and Classifying Mental Disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). Psychological Science in the Public Interest, 18(2), 72-145. https://doi.org/10.1177/1529100617727266
12. American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (DSM-5). Washington, D.C.:American Psychiatric Association.
13. Rana, M., & Bhushan, M. (2023). Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimed Tools Appl 82, 26731–26769. https://doi.org/10.1007/s11042-022-14305-w
14. Iyortsuun, N. K., Kim, S. H., Jhon, M., Yang, H. J., & Pant, S. (2023). A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. Healthcare (Basel, Switzerland), 11(3), 285. https://doi.org/10.3390/healthcare11030285
15. Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. Cureus, 15(4), e37432. https://doi.org/10.7759/cureus.37432
16. Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38.
17. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems, Volume 3, Pages 121-154, ISSN 2667-3452.
18. Rasool, A.; Shahzad, MA.; Aslam, H.; Chan, V. Emotion-Aware Response Generation Using Affect-Enriched Embeddings with LLMs. arXiv 2024, arXiv:2410.01306v1.
19. ChatGPT. [Apr; 2024]. 2024. https://chat.openai.com/
20. Claude. (2024). Claude (May version) [Chatbot].
21. 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 Intelligent Systems, 38(2), 15–23. doi:10.1109/MIS.2023.3254179
22. Pandya, A., Lodha, P., & Ganatra, A. (2024). Is ChatGPT ready to change mental healthcare? Challenges and considerations: a reality-check. Frontiers in Human Dynamics, 5.

doi:10.3389/fhumd.2023.1289255
23. Giadikiaroglou, Panagiotis, et al. "Puzzle solving using reasoning of large language models: A survey." *arXiv preprint arXiv:2402.11291* (2024).
24. Groza, Adrian. "Measuring reasoning capabilities of ChatGPT." *arXiv preprint arXiv:2310.05993* (2023).
25. Bologna, C. (2024, December 19). What is Aphasia?. https://aphasia.org/what-is-aphasia/
26. American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.). https://doi.org/10.1176/appi.books.9780890425787
27. MacWhinney, B. (2019). Understanding spoken language through TalkBank. Behavior Research Methods. doi:10.3758/s13428-018-1174-9
28. MacWhinney, B., Fromm, D., Forbes, M. & Holland, A. (2011). AphasiaBank: Methods for studying discourse. Aphasiology, 25,1286-1307.
29. DeVault, D & Georgilia, K & Artstein, R & Morbini, Fabrizio & Traum, David & Scherer, Stefan & Rizzo, Albert & Morency, L. (2013). Verbal indicators of psychological distress in interactive dialogue with a virtual human. Proceedings of SigDial.
30. Dinkel, H., Wu, M., & Yu, K. (2020). Text-based depression detection on sparse data. arXiv. https://arxiv.org/abs/1904.05154
31. Agrawal, A., & Mishra, H. (2024). Enhancing depression detection in clinical interviews: Integration of fused BERT-BiLSTM model and XGBoost. Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence (ICCAI), Bali Island, Indonesia. https://doi.org/10.1145/3669754.3669780
32. Cong, Yan & Lee, Jiyeon & Lacroix, Arianna. (2024). Leveraging pre-trained large language models for aphasia detection in English and Chinese speakers. 10.18653/v1/2024.clinicalnlp-1.20.
33. K. -C. Chu, Y. -J. Chiu and J. H. B. Masud, "Applying machine learning to language problem analysis," 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI), San Jose, CA, USA, 2024, pp. 200-203, doi: 10.1109/IRI62200.2024.00050.
34. Kertesz A. (2007). Western Aphasia Battery–Revised. San Antonio, TX: The Psychological Corporation.
35. Baron-Cohen, S., Hoekstra, R. A., Knickmeyer, R., & Wheelwright, S. (2006). The Autism-Spectrum Quotient (AQ)--adolescent version. Journal of autism and developmental disorders, 36(3), 343–350. https://doi.org/10.1007/s10803-006-0073-6
36. Burns, D. D. (1980). Feeling good: the new mood therapy. New York, Morrow.
37. Washington, P. (2024) "A Perspective on Crowdsourcing and Human-in-the-Loop Workflows in Precision Health. Journal of Medical Internet Research 26: e51138.
38. Washington, P., & Wall, D. P. (2023). A review of and roadmap for data science and machine learning for the neuropsychiatric phenotype of autism. Annual review of biomedical data science 6(1), 211-228.
39. Tariq, Q., Daniels, J., Schwartz, J. N., Washington, P., Kalantarian, H., & Wall, D. P. (2018). Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS medicine, 15(11), e1002705. https://doi.org/10.1371/journal.pmed.1002705
40. Washington, P., Tariq, Q., Leblanc, E. et al. (2021). Crowdsourced privacy-preserved feature tagging of short home videos for machine learning ASD detection. Scientific Reports 11, 7620. https://doi.org/10.1038/s41598-021-87059-4.
41. Burdisso, S., Reyes-Ramírez, E., Villatoro-Tello, E., Sánchez-Vega, F., López-Monroy, P., & Motlicek, P. (2024). DAIC-WOZ: On the Validity of Using the Therapist’s prompts in Automatic Depression Detection from Clinical Interviews. https://arxiv.org/abs/2404.14463
42. Chen Zhuang, Deng Jiawen, Zhou Jinfeng, Wu Jin- cenzi, Qian Tieyun, and Minlie Huang.

(2024). Depression detection in clinical interviews with LLM- empowered structural element graph. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

43. Ramesh, V., & Assaf, R. (2021). Detecting Autism Spectrum Disorders with Machine Learning Models Using Speech Transcripts. ArXiv, abs/2110.03281.
44. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.
45. Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. arXiv preprint arXiv:2303.17651.
46. Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. arXiv. https://arxiv.org/abs/2307.05300