Leveraging Large Language Models and Traditional Machine Learning Ensembles for ADHD Detection from Narrative Transcripts

Yuxin Zhu*
Yuting Guo*
yuxin.zhu@emory.edu
yuting.guo@emory.edu
Emory University
Atlanta, Georgia, USA

Noah Marchuck
noah.marchuck@emory.edu
Emory University
Atlanta, Georgia, USA

Abeed Sarker
abeed.sarker@emory.edu
Emory University
Atlanta, Georgia, USA

Yun Wang
yun.wang2@emory.edu
Emory University
Atlanta, Georgia, USA

Abstract

Despite rapid advances in large language models (LLMs), their integration with traditional supervised machine learning (ML) techniques that have proven applicability to medical data remains underexplored. This is particularly true for psychiatric applications, where narrative data often exhibit nuanced linguistic and contextual complexity, and can benefit from the combination of multiple models with differing characteristics. Prior research in clinical natural language processing (NLP) has primarily focused on fine-tuning transformer-based models or building domain-specific LLMs, but hybrid approaches that combine deep contextual models with classical ML algorithms have been underexplored. In this study, we introduce an ensemble framework for automatically classifying Attention-Deficit/Hyperactivity Disorder (ADHD) diagnosis (binary) using narrative transcripts. Our approach integrates three complementary models: LLaMA3, an open-source LLM that captures long-range semantic structure; RoBERTa, a pre-trained transformer model fine-tuned on labeled clinical narratives; and a Support Vector Machine (SVM) classifier trained using TF-IDF-based lexical features. These models are aggregated through a majority voting mechanism to enhance predictive robustness. The dataset includes 441 instances, including 352 for training and 89 for validation. Empirical results show that the ensemble outperforms individual models, achieving an F_1 score of 0.71 (95% CI: [0.60-0.80]). Compared to the best-performing individual model (SVM), the ensemble improved recall while maintaining competitive precision. This indicates the strong sensitivity of the ensemble in identifying ADHD-related linguistic cues. These findings demonstrate the promise of hybrid architectures that leverage the semantic richness of LLMs alongside the interpretability and pattern recognition capabilities of traditional supervised ML, offering a new direction for robust and generalizable psychiatric text classification.

Keywords

Large Language Models, Text Classification, Attention-Deficit/Hyperactivity Disorder, Machine Learning, Electronic Health Records, Natural Language Processing

ACM Reference Format:
Yuxin Zhu, Yuting Guo, Noah Marchuck, Abeed Sarker, and Yun Wang. 2018. Leveraging Large Language Models and Traditional Machine Learning Ensembles for ADHD Detection from Narrative Transcripts. In . ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Attention-Deficit/Hyperactivity Disorder (ADHD) is a highly heterogeneous neurodevelopmental condition, presenting diverse etiologies, clinical profiles, and comorbidities [20]. No single risk factor or biomarker conclusively accounts for ADHD's onset; instead, multiple genetic, environmental, and neurodevelopmental factors interplay, leading to varied symptom manifestations in different individuals. Such heterogeneity poses a fundamental challenge for diagnosis. Clinicians rely on behavioral assessments and patient history, but ADHD symptoms (inattention, hyperactivity, impulsivity) often overlap with other disorders (mood, anxiety, learning disorders), making differential diagnosis difficult [10]. There is no single test or "gold-standard" measure to diagnose ADHD in practice—diagnosis typically requires extensive interviews, standardized rating scales from parents and teachers, and the exclusion of alternative explanations [10]. Such comprehensive evaluations are time-consuming and subjective, leading to variability in diagnostic outcomes. In particular, subjectivity and informant bias can affect clinical assessments; parent and teacher ratings may differ widely, and self-reports are prone to error. This subjectivity can

*Both authors contributed equally to this research.

Conference'17, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

yield inconsistent diagnoses, highlighting the need for more objective, reproducible diagnostic aids.

Neuroimaging has been explored as an objective route to ADHD diagnosis, but with limited success. Over two decades of MRI-based studies have revealed some group-level differences in brain structure and function, yet no reliable imaging biomarker has emerged for individual diagnosis. The neuroimaging literature is voluminous and often inconclusive. Meta-analyses, for instance, of ADHD MRI studies frequently find inconsistent or non-overlapping results across samples [17, 24]. Variation in normal brain development, ADHD's own heterogeneity, and small effect sizes mean that MRI findings have not translated into clinically useful tests [19]. In summary, traditional diagnostic methods—clinical evaluation and neuroimaging—are limited by subjectivity, heterogeneity of presentations, and lack of definitive biomarkers. This creates a strong rationale for exploring other computational techniques that can enhance diagnostic accuracy by integrating multi-modal data and reducing human bias.

Recent years have seen growing interest in applying machine learning and data-driven methods to improve ADHD identification and prognosis. Conventional machine learning on clinical datasets has achieved moderate success; for example, decision tree classifiers on neuropsychological and demographic data attained about 75.03% accuracy in distinguishing ADHD cases from non-ADHD controls [7], providing interpretable rules to assist clinicians. Similarly, support vector machines and random forests have been applied to behavioral questionnaires and social media text. one study using a random forest on ADHD-related Reddit posts achieved 81% accuracy in identifying users with ADHD traits [2]. These approaches illustrate the promise of algorithmic classifiers but often remain dataset-specific and can struggle with generalizability due to limited feature scopes.

Natural language processing (NLP) offers another avenue, especially given that clinical text (doctor's notes, psychological reports) and patient self-descriptions contain rich information about ADHD symptoms. NLP techniques have previously been used to detect signs of ADHD in unstructured text. For instance, Malvika et al. applied a transformer-based model (BioClinicalBERT) to electronic health records and achieved an F_1 score of 0.78 in identifying ADHD cases [18], demonstrating the utility of textual data in diagnosis. Likewise, analysis of social media and patient narratives has proven fruitful: fine-tuned transformer models like RoBERTa [16] have reached ~76% accuracy in classifying ADHD-related posts from online forums [15], and SVM classifiers using linguistic and memory-related features attained $F_1 \approx 0.77$ in distinguishing ADHD in personal essays [6]. These advances demonstrate that textual markers of ADHD, such as patterns of language usage, and descriptions of attention difficulties, can be learned by machine learning models, providing a scalable complement to traditional assessments. Notably, researchers are also prioritizing explainability in these models. Kim et al. developed an explainability-enhanced classifier on psychological test reports, which not only achieved high accuracy (~92%) in separating ADHD from intellectual disability, but could also highlight textual evidence (n-gram features) to justify its predictions [14]. This ability to provide evidence-based insights is crucial for physician trust in AI-assisted diagnosis.

In parallel, large language models (LLMs), such as GPT series [5] developed by OpenAI and LLaMA [1] developed by Meta, have demonstrated extraordinary capabilities in understanding and generating human-like text. These models, trained on massive corpora, have been shown to exhibit cognitive-like competencies. They can perform tasks requiring reasoning, memory, and attention—key domains affected in ADHD. Berrezueta-Guzman et al. explored the use of ChatGPT (GPT-3.5) as a conversational agent to support ADHD therapy, finding that experts rated it highly in empathy and adaptability during simulated therapy sessions. Their study highlights that LLMs can capture subtle aspects of communication and patient interaction, reinforcing the idea that such models understand context relevant to ADHD [3]. Another line of research evaluates LLMs on mental health prediction tasks via zero-shot or few-shot prompts. Xu et al. evaluated models like GPT-4, Alpaca, and FLAN-T5 on tasks including detecting ADHD from text, and found zero-shot LLM performance to be promising yet below specialized models. Notably, after fine tuning these LLMs on mental health data, the performance jumped significantly—their fine-tuned "Mental-Alpaca" outperformed even much larger base models by over 10% in balanced accuracy [22]. This reveals that while LLMs have general knowledge (and potentially the ability to recognize ADHD-related patterns learned from text during pretraining), they may need domain-specific tuning to reach diagnostic accuracy. Fine-tuning endows them with jargon understanding and emphasis on the subtle linguistic cues of ADHD, whereas in a zero-shot setting, they might miss context-specific details. In contrast, zero-shot LLMs bring the advantage of flexibility—they can be deployed without task-specific training data, an appealing trait when labeled data is scarce or costly. Beyond direct diagnosis, LLMs are being integrated with healthcare data pipelines. For example, ensemble approaches with multiple LLMs have been used to improve medical question-answering systems. Xiao et al. introduced an LLM ensemble that uses a weighted majority voting among different LLMs to answer medical queries, achieving higher accuracy than any single model by reducing variance and bias [23].

In the broader mental health domain, the integration of LLMs with more traditional analytical models, like supervised machine learning classifiers, remains a relatively open frontier. Studies have repeatedly demonstrated that in the presence of annotated data for training, supervised models outperform zero- or few-shot LLMs relying on in-context learning [21]. While generative LLMs like LLaMA3 have excellent capabilities in understanding nuances of complex, unstructured texts, in-context learning strategies have limitations on how many labeled examples can be provided to the LLM to guide its decision making. This is because the context window sizes of LLMs are limited (e.g., LLaMA3 70B has a context window size of 8000 tokens). At the same time, the large number of parameters in LLMs require very large annotated datasets for meaningful fine-tuning. Thus, traditional supervised machine learning models often have a relative advantage over LLMs when annotated data is available for training. There is, however, little past work attempting to combine the complementary capabilities of LLMs and traditional machine learning models.

In this paper, we describe an ensemble-based ADHD classification model to address this gap. Specifically, we model the ADHD diagnosis as a binary classification task and attempt to solve it by blending state-of-the-art NLP—LLaMA3[1] and RoBERTa, with established machine learning techniques—Support Vector Machine (SVM). This ensemble is, to our knowledge, the first effort to incorporate an LLM with a fine-tuned transformer-based model and a traditional machine learning model for a psychiatric diagnosis task. The specific contributions of this paper are as follows:

- We designed and refined a prompt that specifies the characteristics to consider when deciding if a transcript should be labeled as ADHD or not.
- We developed an LLM (LLaMA3-70B), a transformer-based model (RoBERTa) with supervised learning, and a traditional machine learning model (SVM) for the automatic classification of ADHD cases based on the narrative transcripts.
- We developed an an effective ensemble classification framework that combined the individual predictions under the majority voting scheme.
- We conducted an analysis of the performance and errors for the classifiers which can provide insights for potential future research directions.

2 Materials and Methods

2.1 Data Collection and Narrative Elicitation

Our work utilized a subset of the Healthy Brain Network (HBN) dataset provided by the Child Mind Institute. Following a Data Usage Agreement with Child Mind Institute, we retrieved the verbatim transcripts and clinical diagnosis data for ADHD. The latter is the final diagnosis given by the clinician after administering the KSADS and considering other data and interactions provided as part of participation into the HBN study. Each of the 441 youths (ages 5–21) in our dataset watched an emotionally evocative short animated film ("The Present") during functional MRI scanning; immediately upon exiting the scanner, they completed a structured post-scan interview. A set of open-ended questions was used to elicit narrative recall (e.g., describing the sequence of events), emotional interpretation (e.g., identifying characters' feelings), and perspective-taking (e.g., explaining characters' motivations). These narrative responses formed the transcripts analyzed by our classification models. We chose the transcript data as our target model input because naturalistic narrative language may serve as an alternative, complementary data source that can ecologically assess ADHD. There is growing evidence that children with ADHD exhibit subtle but systematic differences in storytelling—for example, producing narratives that are less coherent, more error-prone, and less richly detailed than those of their peers [13]. In the cases of transcripts from participants with ADHD, responses often lacked direct answers to the interviewer's questions or involved redirecting the conversation by asking questions in return. As shown in the interview excerpt (2) below, when asked to describe the events of a movie, the participant provided a fragmented and disjointed recount. Additionally, the participant frequently expressed confusion or disinterest when asked to elaborate, such as when responding 'I don't know' or shifting the focus of the conversation back to the interviewer. By analyzing such patterns across the entire dataset, our study aims to develop a reliable approach for identifying ADHD from narrative transcripts using three classification models and their ensemble model. Each model approach brings a unique inductive bias to the task of ADHD detection, which is detailed in the following section. The comprehensive experiment pipeline is documented by Figure 1.

[1] We used the LLaMA3-70B model, which we refer to as LLaMA3 for convenience.

2.2 Classification Models

We split the dataset into a training, development, and test set with a ratio of 60/20/20. All classification models were trained using the same training and development set and evaluated on the same test set. By holding the input data constant across models, we ensure that any improvement from combining models is due to their complementary modeling strengths rather than differences in data. Texts were preprocessed by separating interviewer and interviewee texts. For the LLM, the training set was used for testing and optimizing prompts prior to execution on the test set. The dataset maintained a balanced distribution between participants diagnosed with ADHD and non-ADHD (stratified sampling), thus ensuring consistency in class representation throughout all experimental stages. The detailed data statistics are shown in Table 2.

2.2.1 *Large Language Model: LLaMA3.*

We employed a classification methodology utilizing the large language model LLaMA3. This approach leveraged the model's intrinsic capability to interpret narrative content without specific ADHD-focused fine-tuning. The model operate through a mechanism called prompting, where users provide natural language input (known as a prompt) to guide the model's behavior. The prompt frames the task (e.g., classification, summarization, question answering), and the model generates a response based on the contextual information and patterns it has learned during training. The quality and structure of the prompt significantly influence the relevance and accuracy of the output. Each transcript was interpolated into a fixed prompt template (see Figure 3), which instructs the model to adopt the role of a psychiatrist specializing in DSM-5 ADHD diagnosis. Because LLaMA3 supports up to 8,000 tokens of context, no additional segmentation was required for our average transcript length.

Our initial prompt was drafted to mirror the DSM-5 diagnostic criteria for inattention and hyperactivity-impulsivity. We then performed three iterative refinement cycles on the development set. To maximize diagnostic performance in terms of accuracy, precision, recall, and F_1 score, in each cycle, we (1) ran the full development set through the model, (2) logged false positives and false negatives, (3) conducted qualitative error analysis to identify ambiguous or over-broad language in the prompt, and (4) revised the instructions to sharpen symptom definitions and output formatting (for example, explicitly instructing the model to disregard interviewer questions and to respond with "YES." or "NO." at the start of its answer).

For each test-set transcript, we supplied the finalized prompt plus narrative in a single API call. The model's first token was parsed as the binary label (YES = ADHD; NO = non-ADHD), and the subsequent text served as explanatory justification. Decisions were recorded without any post-hoc thresholding or ensembling, isolating the LLaMA3 contribution to our overall framework.

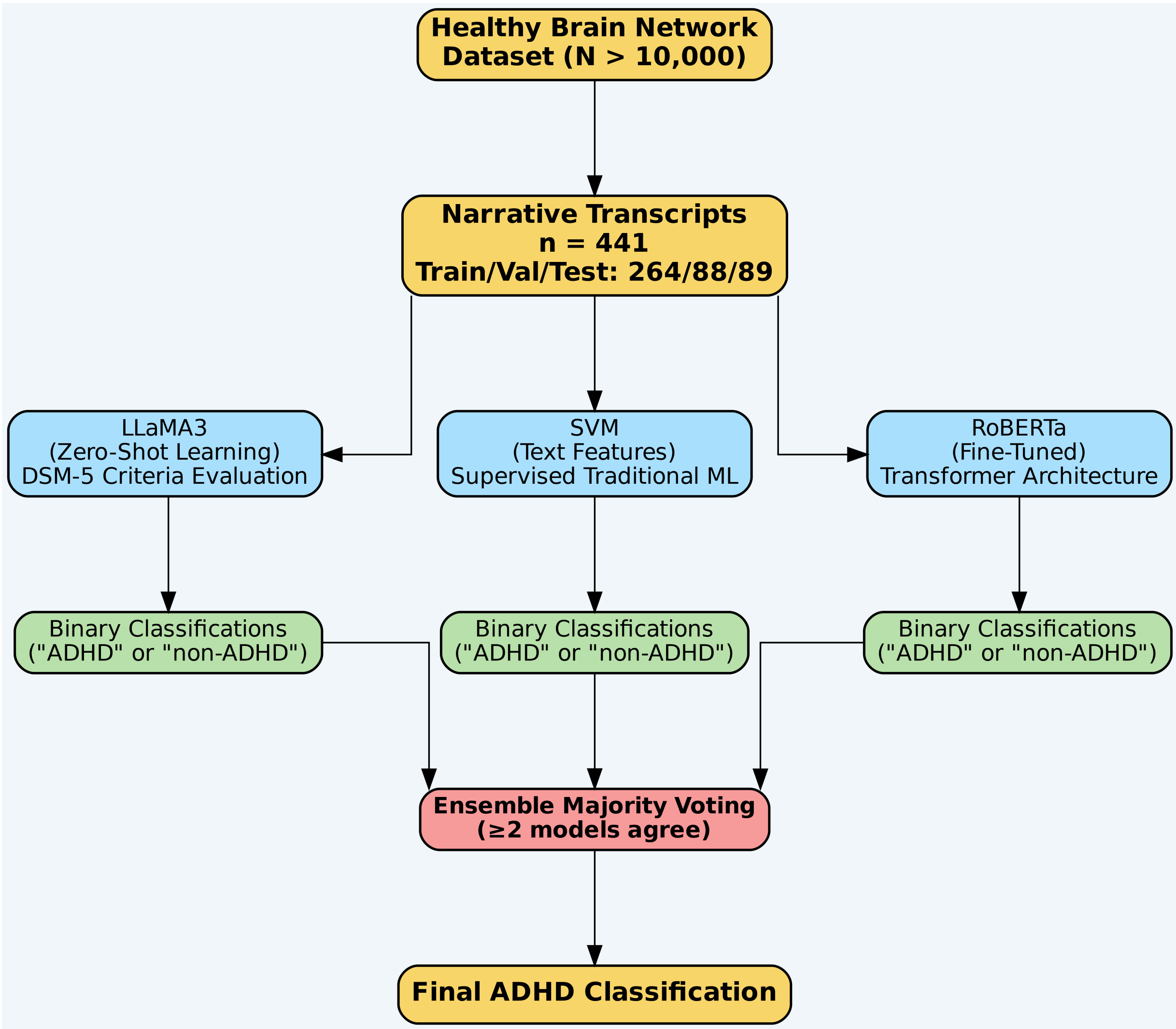


Figure 1: End-to-end ensemble-based narrative classification pipeline. Narrative transcripts elicited immediately after fMRI scanning ("post-scan interviews") undergo preprocessing, including tokenization, TF-IDF vectorization, engineered feature extraction (e.g., response length, question length), and isolation of interviewee texts, before being fed into three complementary classifiers: (1) LLaMA3 via optimized prompt engineering; (2) a fine-tuned RoBERTa transformer; and (3) a support vector machine leveraging both TF-IDF lexical features and additional engineered metrics. Each model independently generates an ADHD vs. non-ADHD prediction, and these are combined under a majority-voting rule (narratives are labeled ADHD if at least two models concur) to produce the final diagnostic classification. Directed arrows denote the flow of data through each module, illustrating the modular architecture of the ensemble framework.

2.2.2 Transformer-Based Model: RoBERTa. RoBERTa is a widely used transformer-based model pre-trained on large English corpora. The model was selected for its proven effectiveness in tasks involving contextual language understanding, and its superior performance reported in classification benchmarking studies [12]. The model takes a text sequence as input, which is first tokenized into word pieces. Each word piece is then encoded into a dense vector. The vector corresponding to the first token is used to represent the entire sequence and is passed through a fully connected layer followed by a sigmoid activation function. The output is a two-dimensional vector representing the predictive scores for the ADHD and non-ADHD classes. The model was explicitly fine-tuned in participant-generated narratives, excluding any interviewer questions to isolate participant-driven linguistic and cognitive patterns. Hyperparameter tuning was performed through iterative experiments by varying the learning rate $\{1\times10^{-5}, 2\times10^{-5}, 4\times10^{-5}\}$ and the number of training epochs $\{10, 15, 20\}$. Model performance was evaluated on the validation set, and the optimal configuration was selected based on validation accuracy. The batch size and maximum sequence length were empirically set to 32 and 512, respectively.

Interviewer: Alright, I hope you enjoyed the last movie. Have you seen it before?

Interviewee: Yeah.

Interviewer: Can you tell me what happened in the movie? And try to tell a little story. Remember that stories have a beginning, things happen, and then some kind of an end.

Interviewee: The kid that was playing a video game, and that the kid was playing a video game. And his mom gave him a package, and when he opened it he saw a puppy. He got a little happy and then he just said "Get lost."

Interviewer: Then what happened?

Interviewee: Then, he actually dropped the ball. And he started going to play outside with the dog.

Interviewer: Do you remember anything else that happened in the movie? No? What are some of the things you liked about the movie?

Interviewee: I don't know.

Interviewer: What did you like?

Interviewee: Why is that camera watching me? Despicable Me.

Interviewer: No, what'd you like about the dog movie?

Interviewee: I don't know. I don't know.

Figure 2: Example of narrative data from a post-scan interview illustrating ADHD-related response patterns. The participant's answers are fragmented and lack coherence, with shifts in focus and expressions of confusion (e.g., "I don't know"). These behaviors—difficulty staying on topic and providing detailed responses—are characteristic of ADHD and serve as key indicators in the classification process.

One major limitation of the RoBERTa model is its maximum input length of 512 tokens, which is often insufficient for processing the narratives that exceed this threshold. To address this, we implemented a sliding window approach with a window size of 512 tokens to divide long narratives into overlapping segments. Each segment was treated as an independent input to the model, allowing for separate predictions. The final prediction for the entire note was then determined by majority voting across the predictions of all segments.

2.2.3 Machine Learning Model: SVM. Support vector machines (SVMs) [8] are well suited to problems with very high-dimensional feature spaces, a characteristic that has underpinned their strong performance in text classification tasks and motivated their selection for the present study. To convert narratives into feature vectors, we applied term frequency–inverse document frequency (TF-IDF) weighting to n-gram representations. In this context, an n-gram denotes any contiguous sequence of n tokens, and our experiments incorporated unigrams (n = 1) through four-grams (n = 4). We retained the 1,000 most common across the training corpus to form our vocabulary. To explore how basic preprocessing might alter downstream performance, we ran each experiment twice under one alternate setting for character normalization: once converting all alphabetic characters to lowercase before tokenization (*lowercase = True*) and once leaving original casing intact (*lowercase = False*). TF-IDF is a statistical measure designed to reflect how important a given term is within a document or corpus. The term frequency (TF) component counts how often each term (i.e., each n-gram) appears in an individual transcript, while the inverse document frequency (IDF) component down-weights terms that occur broadly across the entire training set, thereby emphasizing those that are more distinctive. As a result, the TF-IDF vectors produced assign higher weights to n-grams that are unique to particular documents and lower weights to those that are uniformly distributed throughout the corpus [11].

You are a psychiatrist that is specialized in the diagnosis of ADHD.
Read the transcript of a conversation with an interviewed child, who just finished watching an animated emotionally evocative four-minute film, entitled "The Present", carefully:
{{Transcript Content}}
Please assess the child on the following diagnostic criteria A and B for Attention-Deficit/Hyperactivity Disorder:
A. A persistent pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development, as characterized by (1) and/or (2):
1. Inattention: Six (or more) of the following symptoms have persisted to a degree that is inconsistent with developmental level and that negatively impacts directly on social and academic/occupational activities:
Note: The symptoms are not solely a manifestation of oppositional behavior, defiance, hostility, or failure to understand tasks or instructions.
a. Often fails to give close attention to details or makes careless mistakes in schoolwork, at work, or during other activities (e.g., overlooks or misses details, work is inaccurate).
b. Often has difficulty sustaining attention in tasks or play activities (e.g., has difficulty remaining focused during lectures, conversations, or lengthy reading).
c. Often does not seem to listen when spoken to directly (e.g., mind seems elsewhere, even in the absence of any obvious distraction).
d. Often does not follow through on instructions and fails to finish schoolwork, chores, or duties in the workplace (e.g., starts tasks but quickly loses focus and is easily sidetracked).
e. Often has difficulty organizing tasks and activities (e.g., difficulty managing sequential tasks; difficulty keeping materials and belongings in order; messy, disorganized work; has poor time management; fails to meet deadlines).
f. Often avoids, dislikes, or is reluctant to engage in tasks that require sustained mental effort (e.g., schoolwork or homework; for older adolescents and adults, preparing reports, completing forms, reviewing lengthy papers).
g. Often loses things necessary for tasks or activities (e.g., school materials, pencils, books, tools, wallets, keys, paperwork, eyeglasses, mobile telephones).
h. Is often easily distracted by extraneous stimuli (for older adolescents and adults, may include unrelated thoughts).
i. Is often forgetful in daily activities (e.g., doing chores, running errands; for older adolescents and adults, returning calls, paying bills, keeping appointments).
2. Hyperactivity and impulsivity: Six (or more) of the following symptoms persisted to a degree that is inconsistent with developmental level and that negatively impacts directly on social and academic/occupational activities:
Note: The symptoms are not solely a manifestation of oppositional behavior, defiance, hostility, or a failure to understand tasks or instructions.
a. Often fidgets with or taps hands or feet or squirms in seat.
b. Often leaves seat in situations when remaining seated is expected (e.g., leaves his or her place in the classroom, in the office or other workplace, or in other situations that require remaining in place).
c. Often runs about or climbs in situations where it is inappropriate. (Note: In adolescents or adults, may be limited to feeling restless.)
d. Often unable to play or engage in leisure activities quietly.
e.Is often "on the go," acting as if "driven by a motor" (e.g., is unable to be or uncomfortable being still for extended time, as in restaurants, meetings; may be experienced by others as being restless or difficult to keep up with).
f.Often talks excessively.
g.Often blurts out an answer before a question has been completed (e.g., completes people's sentences; cannot wait for turn in conversation).
h.Often has difficulty waiting his or her turn (e.g., while wating).
i.Often interrupts or intrudes on others (e.g., butts into conversations, or activities; may start using other people's things without asking or receiving permission; for adolescents and adults, may intrude into or take over what others are doing).
B. The symptoms do not occur exclusively during the course of schizophrenia or another psychotic disorder and are not better explained by another mental disorder (e.g., mood disorder, anxiety disorder, dissociative disorder, personality disorder, substance intoxication or withdrawal).
Upon completion of the assessment, you must answer whether the child that was being interviewed has ADHD followed by a detailed justification.
Your answer should be formatted as
'''
YES/NO.
My reasons are ...
'''

Figure 3: The full final prompt used in this study.

Table 1: The statistics for the training, development, and test sets.

Dataset	Size	ADHD	non-ADHD
Train	264	134 (50.76%)	130 (49.24%)
Dev	88	45 (51.14%)	43 (48.86%)
Test	89	45 (50.56%)	44 (49.44%)

Table 2: The accuracy, precision, recall, and F_1 score with 95% confidence intervals (CIs) of LLaMA3, RoBERTa, SVM and an ensemble model using majority voting (MV) on the entire test set.

Model	Accuracy	Precision	Recall	F_1 Score (95% CI)
LLaMA3	0.56	0.54	0.87	0.67 (0.57–0.76)
RoBERTa	0.61	0.57	0.87	0.69 (0.58–0.78)
SVM	0.64	0.62	0.75	0.68 (0.57–0.77)
Ensemble (MV)	0.63	0.59	0.91	0.71 (0.60–0.80)

Recognizing potential feature limitations, we introduced supplementary engineered features calculated directly from narrative transcripts. We conducted experiments under two configurations:

(1) TF–IDF only: using the 1,000 highest-weighted n-gram features derived from TF–IDF.
(2) TF–IDF + engineered features: augmenting the 1,000 TF–IDF features with the following transcript-based metrics:
    - Mean interviewee response length (mean word count per response), hypothesized to capture the verbosity and potential attention-related difficulties such as shorter and less diverse vocabulary usage by some ADHD participants [4].
    - Total number of interviewee responses, anticipated to reflect narrative fragmentation or coherence related to cognitive control deficits.
    - Mean interviewer question length (mean word count per question), included to normalize and account for

interviewer influence on participant narrative length and detail.

Furthermore, a customized tokenization pattern that captures both lexical elements and punctuation was applied to enhance the richness of textual representation, preserving critical linguistic nuances (e.g., pauses, interruptions, emphatic expressions) that may differentiate ADHD and non-ADHD narratives

We performed grid search over the training data to find the best regularization parameter

$$C \in \{2, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024, 2048\}$$

and the kernel type (linear and radial basis function (*rbf*)). Optimal performance, evaluated through cross-validation accuracy and F_1 score, was achieved using a radial basis function kernel and a regularization parameter $C = 1024$. In this configuration, the preprocessing pipeline preserved the original character casing and incorporated the newly engineered features, both of which contributed materially to the observed performance gains.

2.2.4 Ensemble Model. The generative LLM can apply its vast pretrained knowledge to pick up on subtle discourse indicators of inattention or impulsivity that might be reflected by disorganized storytelling or missing plot details, but it may also produce false positives by overgeneralizing. The fine-tuned RoBERTa model learns specific linguistic patterns of ADHD vs. non-ADHD narratives from the training data. Meanwhile, the SVM offers a simpler, interpretable baseline, using engineered features that can highlight straightforward differences. The rationale for an ensemble is that by combining the three weak learners, we can harness their complementary strengths and mitigate individual weaknesses. In theory, an effective ensemble will flag an ADHD case if any one model is sensitive to its particular linguistic quirks, yet require consensus to declare a positive classification, thus filtering out idiosyncratic errors from any single model.

For each participant transcript i, let

$$\hat{y}_i^{\text{LLM}}, \quad \hat{y}_i^{\text{RoBERTa}}, \quad \hat{y}_i^{\text{SVM}} \in \{0, 1\}$$

denote the binary predictions (1 = ADHD, 0 = non-ADHD) produced by the LLaMA3 prompt, the RoBERTa classifier, and the SVM, respectively. The ensemble decision $\hat{y}_i^{\text{Ens}}$ is then given by:

$$\hat{y}_i^{\text{Ens}} = \begin{cases} 1, & \text{if } \hat{y}_i^{\text{LLM}} + \hat{y}_i^{\text{RoBERTa}} + \hat{y}_i^{\text{SVM}} \geq 2, \\ 0, & \text{otherwise.} \end{cases}$$

This majority voting strategy provides a balanced decision rule: a narrative is labeled ADHD only if at least two of the three classifiers agree. Predictions from RoBERTa and the SVM were obtained via their respective predictions on the held-out test set. LLaMA3 outputs were pre-parsed to extract the first token ("YES" → 1, "NO" → 0) before aggregation. We expect that our ensemble will reduce false negatives (by catching cases one model might miss) while controlling false positives (by overriding spurious alerts from one model with the others' dissent), thereby improving overall diagnostic accuracy.

3 Evaluation Metrics

Model performances were measured by the precision, recall (sensitivity), and F_1 score (harmonic mean of precision and recall) metrics for the ADHD class. The F_1 was used as the primary evaluation metric to balance precision and recall, ensuring that improvements in one did not come at the expense of the other. Bootstrap resampling was used to compute the 95% confidence intervals of F_1 scores [9]. Resampling was performed with replacement ($N = 1000$) over 1000 iterations, and the 2.5^{th} and 97.5^{th} percentile scores were selected as the interval boundaries. We also analyzed the confusion matrix for each classification model to examine their error patterns, including false positives and false negatives, and to better understand how each model performs across the ADHD and non-ADHD classes.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

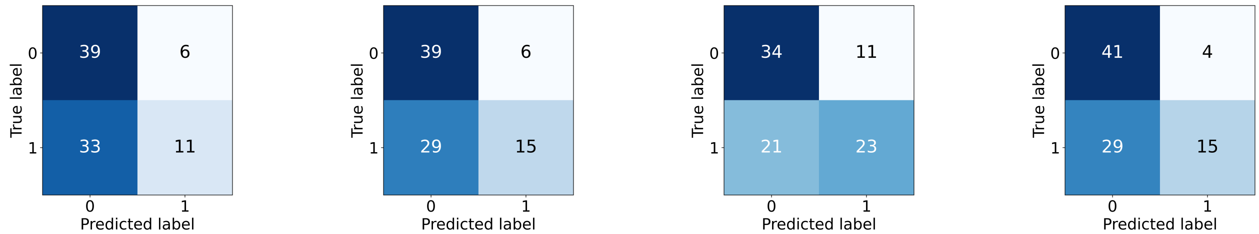
$$F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total \# Instances}} \tag{4}$$

4 Results

Table 2 presents the performance of four models. Among the individual models, SVM achieved the highest accuracy (0.64) and precision (0.62), while RoBERTa and SVM achieved recall of 0.87 and 0.75, respectively. The ensemble model outperformed all others in terms of recall (0.91) and F_1 score (0.71), indicating its effectiveness in identifying positive instances with a better balance between precision and recall. LLaMA3 showed the lowest accuracy (0.56) and precision (0.54), though its recall remained high (0.87), suggesting that it tends to over-predict positive cases. The 95% confidence intervals for the F_1 score further support the robustness of the ensemble approach, with the ensemble achieving the highest upper bound (0.80) and a relatively narrow interval (0.60–0.80), highlighting its stability across runs.

Figure 4 presents the confusion matrices for the four classification models. LLaMA3 demonstrates a high recall for ADHD cases, correctly identifying 11 out of 44 ADHD instances, but misclassifies a large number of true ADHD cases as non-ADHD (33), indicating a strong tendency toward over-predicting the negative class. RoBERTa improves upon this with 15 correct ADHD classifications and fewer false negatives (29), achieving a more balanced performance between sensitivity and specificity. SVM shows the most balanced classification across both classes, with 23 true positives and 21 false negatives for ADHD, and 11 false positives for non-ADHD, reflecting its relatively high precision. The ensemble model achieves the best overall performance by minimizing false positives (only 4) and increasing correct classifications of non-ADHD (41), while maintaining a similar true positive rate (15) to RoBERTa. This confirms the ensemble's advantage in reducing classification bias and improving reliability across both classes. Overall, the confusion matrices support the quantitative findings by highlighting



(a) LLaMA3 **(b) RoBERTa** **(c) SVM** **(d) Ensemble**

Figure 4: The confusion matrices for each individual model and the ensemble model.

each model's classification behavior and reinforcing the ensemble's effectiveness in enhancing both precision and recall.

4.1 Qualitative Error Analysis

To complement the quantitative confusion matrices, we examined a sample of misclassifications from the held-out test set to uncover model-specific failure modes:

- SVM false positives: Two non-ADHD transcripts containing repetitive patterns of confusion ("I don't know") were overweighted by TF–IDF, leading the SVM to mislabel them as ADHD.
- SVM false negatives: In contrast, some ADHD transcripts with more formal, coherent language lacked the distinctive n-grams the SVM relied on, resulting in missed detections.
- RoBERTa errors: RoBERTa's false negatives typically occurred when ADHD participants delivered concise but disfluent narratives (e.g., short, choppy sentences), suggesting the model places undue weight on explicit DSM-5 keywords rather than discourse coherence.
- LLaMA3 errors: LLaMA3's false negatives often involved transcripts where the prompt-engineered criteria focused on symptom keywords over narrative structure, causing it to overlook subtle signs of inattention.

These examples reveal that each model is sensitive to different linguistic cues—n-gram sparsity for the SVM, contextual token patterns for RoBERTa, and prompt formulation for LLaMA3—and point toward future improvements such as incorporating syntactic cohesion metrics or refining prompt definitions to better capture narrative coherence.

5 Discussion

The results presented in Table 2 demonstrate the relative strengths and limitations of each model in the ADHD classification task. The ensemble model, which integrates predictions from LLaMA3, RoBERTa, and SVM through majority voting, achieved the highest F_1 score (0.71) and recall (0.91), indicating its superior ability to identify ADHD cases with both high sensitivity and balanced performance. This suggests that combining diverse model architectures—ranging from large language models to traditional machine learning classifiers—can lead to more robust and generalizable outcomes in psychiatric classification tasks. Among the individual models, SVM achieved the highest accuracy (0.64) and precision (0.62), highlighting its reliability in reducing false positives. In contrast, LLaMA3 showed the lowest precision (0.54) and accuracy (0.56) but maintained a high recall (0.87), implying that while it is effective at capturing true ADHD cases, it may also over-predict positive labels. RoBERTa displayed moderate performance across all metrics, balancing between the high recall of LLaMA3 and the precision of SVM.

Moreover, a closer inspection of the prediction distributions (Table 2 and Figure 4) reveals that the SVM classifier issues a higher proportion of positive (ADHD) labels than both LLaMA3 and RoBERTa. We hypothesize three contributing factors:

1. **Regularization and decision boundary width.** The optimal hyperparameter C=1024 creates a narrow margin that tightly fits the training examples, making the SVM hypersensitive to distinctive n-grams that co-occur with ADHD labels and thus more prone to over-predict positives.
2. **TF–IDF feature weighting.** By design, TF–IDF amplifies rare but highly distinctive n-grams. If such n-grams are noisy or spuriously correlated with ADHD in our 1 000-feature vocabulary, the decision function may tilt toward the positive class.
3. **Residual class imbalance in high-dimensional space.** Although our dataset was stratified to 50/50 ADHD vs. non-ADHD, the effective class balance in the transformed feature space can shift due to sparsity. This can bias the hyperplane toward the denser region of the ADHD samples.

To mitigate this positive bias in future work, we plan to (a) recalibrate the SVM decision threshold on a validation set, (b) introduce class-weighted penalties or cost-sensitive learning, and (c) conduct ablation studies on n-gram ranges and engineered features to identify and remove noisy predictors.

These findings illustrate that while LLMs like LLaMA3 can extract nuanced patterns from unstructured text, they may benefit from complementary methods to enhance precision. The ensemble approach capitalizes on the unique strengths of each model—LLaMA3's language understanding, RoBERTa's fine-tuned classification, and SVM's interpretability and structure-based decision-making—resulting in improved overall performance. Notably, the ensemble's F_1 score confidence interval (0.60–0.80) was both higher and tighter than those of the individual models, further supporting its reliability. This work provides early evidence that integrating LLMs with traditional and transformer-based models can be a promising direction in psychological and psychiatric informatics. Future work could explore more sophisticated ensemble strategies, such as weighted voting or stacking, and assess generalizability across larger and more diverse clinical datasets.

5.1 Limitations

While our ensemble-based approach demonstrates promising results for ADHD classification using narrative transcripts, several limitations should be acknowledged. First, the dataset used in this study is relatively small and may not capture the full variability present in broader clinical or community populations. This limits the generalizability of our findings and raises the potential for overfitting. This limitation also highlights the need for more labeled data to train more robust and customized systems. Second, the narrative data were derived from a specific source and may reflect biases in language use or reporting styles that differ across contexts or demographic groups. Third, while the ensemble model improves overall performance, it relies on majority voting, which does not consider model confidence or weight individual model contributions, potentially limiting its adaptability. Finally, the interpretability of LLMs such as LLaMA3 remains a challenge, making it difficult to fully understand the decision-making process behind individual predictions. Future work should explore larger and more diverse datasets, evaluate model generalization across settings, and investigate more advanced ensemble techniques that incorporate confidence scores or learn optimal weights. In addition, integrating interpretable NLP techniques may help enhance model transparency and clinical trust.

6 Conclusion

In this study, we proposed a reliable ensemble-based framework for classifying ADHD from narrative transcripts, combining a large language model (LLaMA3), a transformer-based model (RoBERTa), and a traditional machine learning classifier (SVM). Our results show that while individual models offer distinct advantages—such as high recall from LLaMA3 and high precision from SVM—the ensemble model consistently outperformed all individual approaches across key evaluation metrics, particularly in terms of F_1 score and recall. This highlights the potential of integrating LLMs with conventional models to enhance diagnostic classification in the psychological domain. Our findings open new avenues for hybrid modeling approaches in mental health applications and underscore the value of model diversity in building robust and interpretable clinical decision support tools. However, further research is needed to validate these findings on larger and more diverse datasets and to explore advanced ensemble techniques and longitudinal predictive modeling.

References

[1] AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] Nizar Alsharif, Mosleh Hmoud Al-Adhaileh, Saleh Nagi Alsubari, and Mohammed Al-Yaari. 2024. ADHD Diagnosis Using Text Features and Predictive Machine Learning and Deep Learning Algorithms. *Journal of Disability Research* 3, 7 (2024). doi:10.57197/JDR-2024-0082

[3] Santiago Berrezueta-Guzman, Mohanad Kandil, María-Luisa Martín-Ruiz, Iván Pau de la Cruz, and Stephan Krusche. 2024. Future of ADHD Care: Evaluating the Efficacy of ChatGPT in Therapy Enhancement. *Healthcare (Basel)* 12, 6 (2024), 683. doi:10.3390/healthcare12060683

[4] ELLEN BIALYSTOK, KORNELIA HAWRYLEWICZ, MELODY WISEHEART, and MAGGIE TOPLAK. 2017. Interaction of bilingualism and Attention-Deficit/Hyperactivity Disorder in young adults. *Bilingualism: Language and Cognition* 20, 3 (2017), 588–601. doi:10.1017/S1366728915000887

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

[6] Florian Raphaël Cafiero, Juan Barrios Rudloff, and Simon Gabay. 2024. Harnessing Linguistic Analysis for ADHD Diagnosis Support: A Stylometric Approach to Self-Defining Memories. In *Proceedings of the Fifth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, Dimitrios Kokkinakis, Kathleen C. Fraser, Charalambos K. Themistocleous, Kristina Lundholm Fors, Athanasios Tsanas, and Fredrik Ohman (Eds.). ELRA and ICCL, Torino, Italia, 87–94. https://aclanthology.org/2024.rapid-1.10/

[7] Tianhua Chen, Ilias Tachmazidis, Sotiris Batsakis, Marios Adamou, Emmanuel Papadakis, and Grigoris Antoniou. 2023. Diagnosing attention-deficit hyperactivity disorder (ADHD) using artificial intelligence: a clinical study in the UK. *Frontiers in Psychiatry* Volume 14 - 2023 (2023). doi:10.3389/fpsyt.2023.1164433

[8] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20 (1995), 273–297.

[9] Bradley Efron. 1979. Another Look at the Jackknife. *The Annals of Statistics* 7 (1979), 1–26. Issue 1.

[10] C Thomas Gualtieri and Lynda G Johnson. 2005. ADHD: Is Objective Diagnosis Possible? *Psychiatry (Edgmont)* 2, 11 (November 2005), 44–53.

[11] Yuting Guo, Mohammed A Al-Garadi, Wendy M Book, Lindsey C Ivey, Fred H Rodriguez III, Cheryl L Raskind-Hood, Chad Robichaux, and Abeed Sarker. 2023. Supervised text classification system detects Fontan patients in electronic records with higher accuracy than ICD codes. *Journal of the American Heart Association* 12, 13 (2023), e030046.

[12] Yuting Guo, Xiangjue Dong, Mohammed Ali Al-Garadi, Abeed Sarker, Cecile Paris, and Diego Mollá Aliod. 2020. Benchmarking of Transformer-Based Pre-Trained Models on Social Media Text Classification Datasets. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, Maria Kim, Daniel Beck, and Meladel Mistica (Eds.). Australasian Language Technology Association, Virtual Workshop, 86–91. https://aclanthology.org/2020.alta-1.10/

[13] Ida Bonnerup Jepsen, Esben Hougaard, Susan Tomczak Matthiesen, and Rikke Lambek. 2022. A Systematic Review and Meta-analysis of Narrative Language Abilities in Children with Attention-Deficit/Hyperactivity Disorder. *Res Child Adolesc Psychopathol* 50, 6 (June 2022), 737–751. doi:10.1007/s10802-021-00871-4

[14] Tong Min Kim, Young-Hoon Kim, Sung-Hee Song, In-Young Choi, Dai-Jin Kim, and Taehoon Ko. 2024. Explainability Enhanced Machine Learning Model for Classifying Intellectual Disability and Attention-Deficit/Hyperactivity Disorder With Psychological Test Reports. *J Korean Med Sci* 40, 11 (November 2024). doi:10.3346/jkms.2025.40.e26

[15] Claire S. Lee, Noelle Lim, and Michael Guerzhoy. 2024. Detecting a Proxy for Potential Comorbid ADHD in People Reporting Anxiety Symptoms from Social Media Data. arXiv:2403.05561 [cs.CY] https://arxiv.org/abs/2403.05561

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *arXiv* (Online). arXiv. Issue 1. doi:10.48550/arXiv.1907.11692

[17] Valeria Parlatini, Takashi Itahashi, Yeji Lee, Siwei Liu, Thuan T. Nguyen, Yuta Y. Aoki, Stephanie J. Forkel, Marco Catani, Katya Rubia, Juan H. Zhou, Declan G. Murphy, and Samuele Cortese. 2023. White matter alterations in Attention-Deficit/Hyperactivity Disorder (ADHD): a systematic review of 129 diffusion imaging studies with meta-analysis. *Mol Psychiatry* 28 (2023), 4098–4123. doi:10.1038/s41380-023-02173-1

[18] Malvika Pillai, Jose Posada, Rebecca M Gardner, Tina Hernandez-Boussard, and Yair Bannett. 2024. Measuring quality-of-care in treatment of young children with attention-deficit/hyperactivity disorder using pre-trained language models. *Journal of the American Medical Informatics Association* 31, 4 (01 2024), 949–957. doi:10.1093/jamia/ocae001 arXiv:https://academic.oup.com/jamia/article-pdf/31/4/949/57148541/ocae001.pdf

[19] Alfredo A. Pulini, Wesley T. Kerr, Sandra K. Loo, and Agatha Lenartowicz. 2019. Classification Accuracy of Neuroimaging Biomarkers in Attention-Deficit/Hyperactivity Disorder: Effects of Sample Size and Circular Analysis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 4, 2 (2019), 108–120. doi:10.1016/j.bpsc.2018.06.003

[20] Laura Reale, Beatrice Bartoli, Massimo Cartabia, Michele Zanetti, Maria Antonella Costantino, Maria Paola Canevini, Cristiano Termine, Maurizio Bonati, and Lombardy ADHD Group. 2017. Comorbidity prevalence and treatment outcome in children and adolescents with ADHD. *Eur Child Adolesc Psychiatry* 26, 12 (December 2017), 1443—1457. doi:10.1007/s00787-017-1005-z

[21] Abeed Sarker, Rui Zhang, Yanshan Wang, Yunyu Xiao, Sudeshna Das, Dalton Schutte, David Oniani, Qianqian Xie, and Hua Xu. 2024. Natural Language

Processing for Digital Health in the Era of Large Language Models. *Yearbook of Medical Informatics* 33, 01 (2024), 229–240. doi:10.1055/s-0044-1800750

[22] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–32. doi:10.1145/3643540

[23] Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. 2023. One LLM is not Enough: Harnessing the Power of Ensemble Learning for Medical Question Answering. *medRxiv* (2023). doi:10.1101/2023.12.21.23300380 arXiv:https://www.medrxiv.org/content/early/2023/12/24/2023.12.21.23300380.full.pdf

[24] Miaomiao Yu, Xinyu Gao, Xiaoyu Niu, Mengzhe Zhang, Zhengui Yang, Shaoqiang Han, Jingliang Cheng, and Yong Zhang. 2023. Meta-analysis of structural and functional alterations of brain in patients with attention-deficit/hyperactivity disorder. *Frontiers in psychiatry* 13 (2023), 1070142.