Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
5(2): 127–153.
DOI: 10.3934/aci.2025009
Received: 13 March 2025
Revised: 16 July 2025
Accepted: 17 July 2025
Published: 22 July 2025

Research article

Large language model enabled synthetic dataset generation for human-AI teaming in mental health assessment

Sai Sanjay Potluri*, **Md Abdullah Al Hafiz Khan and Yong Pei**

College of Computing and Software Engineering, Kennesaw State University, 1100 South Marietta Pkwy SE, Marietta, GA 30060, USA

* **Correspondence:** Email: spotlur4@students.kennesaw.edu.

Academic Editor: Jidong Yang

Abstract: Mental health assessment presents unique challenges in healthcare due to its inherently subjective nature and the scarcity of high-quality training data. This research explores the use of large language models (LLMs) to generate synthetic datasets for Human-AI teaming algorithms, focusing on mental health assessments. We created a diverse dataset that simulates human-AI collaboration scenarios in diagnostic processes. Synthetic data are labeled through an innovative approach that involves two human annotators and three LLMs, using majority voting for consensus-based annotations. The dataset initially achieves a similarity score of 83.1%, which is further improved by considering human factors that influence decision-making, with varying performance across different mental health categories, highlighting the need for targeted improvements in data collection and model architecture. Our approach addresses several key challenges in the field, including the lack of real-world training data, privacy concerns, and the need for diverse training datasets. This study serves as a foundation for future work in this critical area, which could lead to more effective and ethically sound AI-assisted mental health assessment tools.

Keywords: artificial intelligence; machine learning; large-language models; datasets; synthetic data

1. Introduction

Mental health disorders are among the leading causes of disability worldwide, yet their evaluation remains one of the most challenging areas in healthcare due to their subjective nature. Unlike physical illnesses that can often be diagnosed through objective tests, mental health conditions rely heavily on behavioral observations, self-reports, and clinical expertise. This subjectivity introduces variability in diagnoses and complicates efforts to develop machine learning models for automated or semi-automated mental health assessments [10].

Recent advances in large language models have opened new possibilities for generating synthetic data across various domains [7]. In healthcare and mental health specifically, synthetic data generation using LLMs presents opportunities to augment limited real-world data sets while preserving privacy [8]. However, leveraging LLMs to create data sets that capture the nuances of human-AI collaboration in clinical decision making remains an area that has not been explored [4]. The scarcity of high-quality labeled datasets further compounds these challenges. Real-world mental health data are difficult to collect due to privacy concerns and ethical considerations surrounding patient confidentiality [12]. Furthermore, existing data sets often lack diversity in terms of patient demographics, cultural contexts, and symptom presentations, limiting their utility to train robust AI models. This research aims to address these limitations by developing a novel approach to generate synthetic datasets that simulate human-AI collaboration in mental health diagnostics. The novelty of our approach lies in its focus on capturing the dynamics of interdisciplinary collaboration between human experts and AI systems. By integrating human expertise with multiple LLMs during the annotation process, we create a dataset that reflects the complexities of real-world decision making while maintaining ethical safeguards.

The research introduces a novel methodology to generate synthetic mental health assessment scenarios using large language models. By simulating realistic diagnostic conversations, this approach augments existing mental health datasets, which are often limited due to privacy concerns. It captures the variability of real-world scenarios while ensuring privacy, and addressing challenges related to data scarcity and confidentiality in mental health research.

A unique annotation framework is proposed that combines human expertise with AI-generated classifications to improve the quality and reliability of synthetic data. Human annotators guide AI models in labeling the data, ensuring that complex, subjective elements of mental health assessments (e.g., behavioral observations, self-reports) are accurately captured. This collaboration helps generate more robust datasets that reflect real-world clinical decision-making processes.

The research delivers a large-scale synthetic dataset specifically designed to explore Human-AI collaboration in mental health diagnostics. It is enriched with diverse demographic data from patients, cultural contexts, and symptom variations, overcoming limitations in current datasets. The study also explores the ethical and practical implications of using LLMs in healthcare care, emphasizing the need for ethical safeguards, privacy preservation, and the potential for AI to complement human expertise in mental health decision-making.

By generating high-quality synthetic datasets, the research addresses critical challenges such as data scarcity, privacy concerns, and the lack of diversity in existing datasets [8]. The proposed framework enables the exploration of Human-AI collaboration dynamics, fostering the development of intelligent systems that complement human expertise in complex decision making. Beyond mental health, this approach can be adapted to other domains that require interdisciplinary collaboration, such as education and social work. By bridging computational innovation with clinical insight, this work contributes to building ethical, scalable, and privacy-preserving AI systems that enhance societal well-being.

2. Systematic overview

In recent years, the proliferation of digital platforms and social networks has provided an unprecedented opportunity to capture and analyze large-scale data related to mental health. Machine learning and NLP techniques have shown promise in detecting linguistic patterns and indicators of suicidal ideation in various text-based data sources, such as social media posts, online forums, and electronic health records [3]. The goal is to address challenges such as data scarcity, privacy concerns, and the complexity of human-AI collaboration in diagnostic processes. In the following, we provide a detailed explanation of the approaches used, the project pipeline, classifier selection, and technological solutions. The primary approach involves leveraging LLMs to generate synthetic data that simulates real-world mental health diagnostic scenarios, including collecting the prompts on how an individual is feeling in their own words to keep the base data as natural as possible. This is then passed through an annotation framework that classifies the prompts. Upon categorization given by the annotations, the resulting diagnosis is then passed through a majority-voting technique on the five classifications. Resulting in a single categorization from combining human expertise with AI-generated classifications through a consensus-based majority voting mechanism, this will later be used as a measuring parameter while getting the agreement statistics.

The system is designed (as shown in Figure 1) as a complete multistage pipeline to generate, annotate, and analyze synthetic datasets for mental health assessment. It integrates two critical components: generating of text data and annotation framework. Initially, synthetic scenarios are meticulously generated using three advanced large language models: Phi-3.5-mini-instruct, Mistral-8x7b-32768, and Gemma2-9b-it based on carefully crafted prompts that simulate complex, real-world mental health cases. These scenarios are then annotated through a sophisticated hybrid process involving clinical human experts and AI-driven LLM classifications, with a robust consensus-building mechanism ensuring reliable and accurate labels. The final dataset undergoes rigorous analysis using advanced metrics like cosine similarity, entropy, and cognitive load assessment to evaluate consistency, quality, and computational complexity, ultimately providing a comprehensive framework for Human-AI teaming in mental health diagnostics.

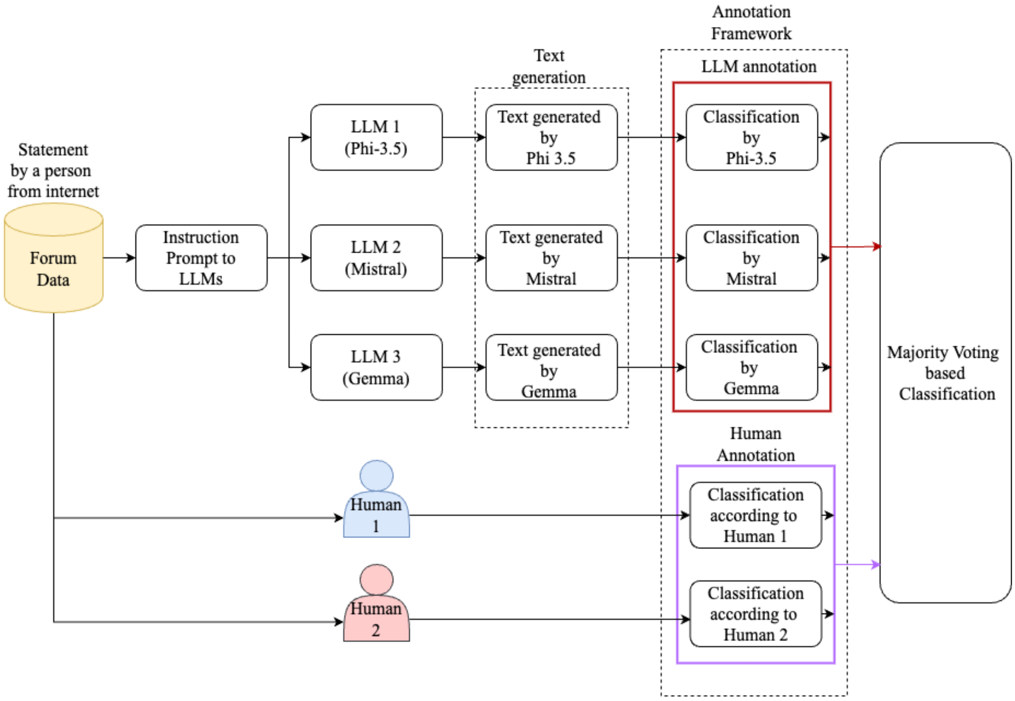
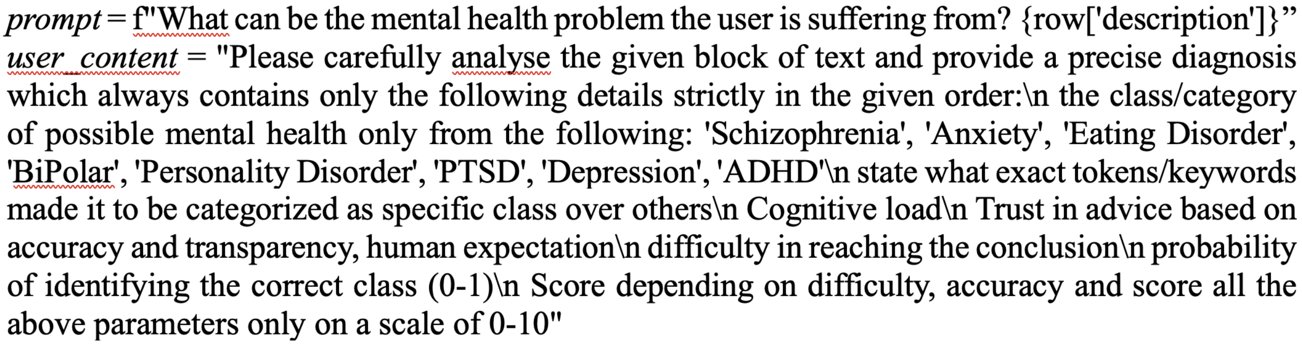

Annotation Framework
Text generation
LLM annotation
Statement by a person from internet
Forum Data
Instruction Prompt to LLMs
LLM 1 (Phi-3.5)
LLM 2 (Mistral)
LLM 3 (Gemma)
Text generated by Phi 3.5
Text generated by Mistral
Text generated by Gemma
Classification by Phi-3.5
Classification by Mistral
Classification by Gemma
Majority Voting based Classification
Human Annotation
Human 1
Human 2
Classification according to Human 1
Classification according to Human 2

Figure 1. Systematic overview of the data generation pipeline.

3. Data generation pipeline

3.1. Prompt engineering for text generation

Prompts were designed to simulate real-world mental health scenarios in eight categories. These synthetic texts generated by the LLMs based on the description given by the user are then saved into the data set initially under the {model}_output column where the model denotes the LLM which generated the output. To ensure that all models generate a similar output so that no one has an advantage over others, the same instruction prompt is passed as shown in Figure 2. The prompt that was given was constructed after multiple trail and errors with careful consideration of noticing how each new instruction added affects the LLM text generation behaviors [1].



prompt = f"What can be the mental health problem the user is suffering from? {row['description']}”
user_content = "Please carefully analyse the given block of text and provide a precise diagnosis which always contains only the following details strictly in the given order:\n the class/category of possible mental health only from the following: 'Schizophrenia', 'Anxiety', 'Eating Disorder', 'BiPolar', 'Personality Disorder', 'PTSD', 'Depression', 'ADHD'\n state what exact tokens/keywords made it to be categorized as specific class over others\n Cognitive load\n Trust in advice based on accuracy and transparency, human expectation\n difficulty in reaching the conclusion\n probability of identifying the correct class (0-1)\n Score depending on difficulty, accuracy and score all the above parameters only on a scale of 0-10"

Figure 2. Prompt passed through LLMs.

Here, the initial command “role”: “system” defines the context for the system which gives a brief overview of the task to be performed. The latter “role”: “user” provides a precise instruction about the expected contents in the output. This can be used to obtain a consistent format of output from the LLMs. This plays a major role in producing a concise output which makes it efficient while extracting the model-based annotation in the subsequent process. But since the information collected initially is noisy and is required to be processed under various layers of text slicing to turn it into training data which can be used to train the model in the long run. The response generated by large language models consists of all the required information starting from possible diagnosis of the mental health condition according to AI to the transparency and accuracy values of the generated text.

The information is then extracted from the generated text using slicing methods done with the help of python’s regular expression library named RegEx. The extracted data resulted from text slicing are saved into an initial dataset resulting a dataset which contains 2,040 entries, each with 12 features, including scenario descriptions, individual annotations, and consensus classifications.

3.2. Model selection

The selected LLMs (Table 1) were chosen for their complementary strengths in instruction follow-up, long-context processing, and task-specific generation, ensuring robustness in synthetic data generation. The selection criteria included their specialization in instruction-following tasks, long-context processing, and task-specific generation. Each model was evaluated for its ability to generate diverse, coherent, and clinically relevant outputs. The models were chosen to ensure robustness and diversity in the dataset, balancing computational efficiency with the complexity required for nuanced mental health diagnostics. The high number of parameters in the selected large language models provides several advantages, mainly related to their ability to process, understand, and generate complex language. The parameters represent the weights learned during training, and increasing their count increases the model's capacity to capture intricate patterns in data, including grammar, syntax, semantics, and context. This allows larger models to generate coherent, contextually relevant text and perform better across diverse tasks. Larger models with more parameters can store more information and nuances from the training data, leading to higher accuracy and improved generalization to unseen input. They are particularly effective for applications requiring deep contextual understanding, such as conversational AI or content generation. Additionally, models with extensive parameters often exhibit flexibility and extensibility, allowing fine-tuning for specific tasks or domains without needing entirely new training. However, this comes at the cost of increased computational demands, memory usage, and latency.

Table 1. A comparison of used large language models, highlighting their key characteristics.

Model	Parameters	Specialization
Phi-3.5-mini-instruct	3.5B	Instruction following
Mistral-8x7b-32768	8x7B	Long-context processing
Gemma2-9b-it	9B	Task-specific generation

3.3. Annotation framework

The annotation framework is a critical component of the methodology, designed to ensure the reliability and precision of the synthetic data set. It integrates human expertise with AI-generated classifications through a multi-stage process. Initially, two clinical experts independently annotated each scenario based on their professional judgment, providing a benchmark for comparison. Simultaneously, three large language models classify the same scenarios, offering predictions and confidence scores. The consensus building algorithm is then used to finalize the annotations. This algorithm uses majority voting to resolve disagreements between annotators, prioritizing human classifications in ambiguous cases. This hybrid approach ensures high-quality labels while leveraging the strengths of both human expertise and AI capabilities.

3.3.1. Human annotation

Human annotation involves the manual labeling of mental health data by understanding the user's emotion in the description. To ensure the precision and relevance of the data set while maintaining diversity, two individuals independently reviewed and classified each synthetic scenario into predefined mental health categories (e.g., anxiety, depression, schizophrenia). This process served as a reference for evaluating the performance of large language models. Human annotators relied on their clinical experience to make informed decisions about each scenario. The annotations were not limited to assigning a single label; annotators were instructed to pay close attention to every single word from the description so that their classifications are concise and shall have a sound reasoning. This approach ensured that the data set captured nuanced and realistic representations of mental health conditions. Human annotation is particularly valuable for addressing ambiguities in the data and validating the outputs of automated systems. However, it is time-consuming and subject to individual biases as represented in Figure 3, which highlights the importance of integrating it with AI-based methods to improve scalability and consistency [11].

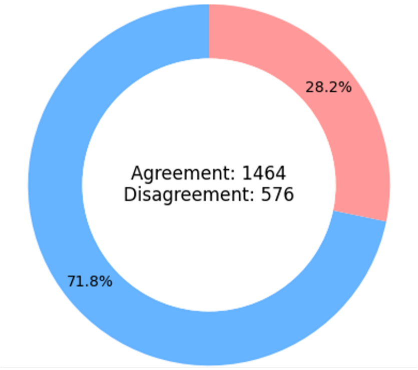


Figure 3. Classification agreements between human annotators.

3.3.2. Annotation by LLMs

LLM classification takes advantage of advanced natural language processing capabilities of large language models to automatically analyze and classify mental health scenarios. In this study, three state-of-the-art LLMs Phi-3.5-mini-instruct, Mistral-8x7b-32768 and Gemma2-9b-it were used to classify synthetic scenarios into mental health categories. Each LLM was prompted with detailed descriptions of patient symptoms and contextual information, which generated classifications along with confidence scores for each prediction. The models were chosen for their complementary strengths: Phi-3.5-mini-instruct excelled in instruction-following tasks, Mistral-8x7b-32768 handled long-context inputs effectively, and Gemma2-9b-it provided nuanced task-specific outputs [9]. The LLMs also offered rationales for their classifications, improving interpretability. These outputs were compared with human annotations to assess alignment and reliability. Although LLMs offer scalability and speed in processing large datasets, they are prone to biases from training data and may struggle with complex or ambiguous scenarios. To mitigate these limitations, a consensus-based annotation framework was used that combines human expertise with LLM predictions to finalize classifications. Each model received a two-part prompt:

System Role: "What can be the mental health problem the user is suffering from?" followed by the scenario description.

User Role: A detailed instruction specifying strict output requirements: "Please carefully analyze the given block of text and provide a precise diagnosis which always contains only the following details strictly in the given order: the class/category of possible mental health only from the following: [Schizophrenia, Anxiety, Eating Disorder, BiPolar, Personality Disorder, PTSD, Depression, ADHD]."

This approach ensured consistent, structured outputs from all LLMs (e.g., "Anxiety"). The models generated classifications alongside confidence scores and rationales, enhancing interpretability. These outputs were compared against human annotations to evaluate alignment and reliability. While LLMs offer scalability, they may inherit biases or struggle with ambiguous cases. To mitigate this, our consensus framework (Section 3.4) integrates human expertise with LLM predictions to finalize classifications. To quantify consensus among LLMs, we categorized agreement levels into three types:

- **Exact Agreement**: All three LLMs assign identical classifications.

- **Partial Agreement**: One or more LLMs concur, while one or more disagrees.
- **No Agreement**: All three LLMs provide divergent classifications.

Figure 4 visualizes these level of agreement across the data set, revealing that LLMs achieved partial agreement in 37.9% of cases (773/2040), primarily due to nuanced differences in the interpretation of symptoms (for example, the overlapping features of *Anxiety* vs. *Depression*). This metric highlights the complexity of mental health categorization and underscores the need for human-AI consensus mechanisms.

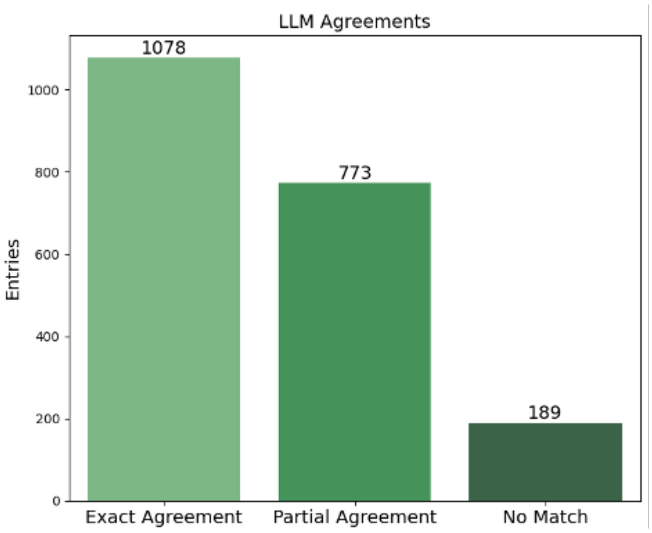


Figure 4. Agreements of different LLMs.

3.4. Majority voting

The majority voting system for mental health classification integrates multiple components, three large language models and two human annotators to provide a robust and reliable decision-making framework from Figure 5. This system ensures a comprehensive evaluation of mental health cases leveraging diverse perspectives and expertise from both AI and humans.

The three LLMs are responsible for generating initial predictions based on their respective strengths. Phi handles instruction follow-up tasks, interprets mental health questions, and provides structured responses. Mistral specializes in analyzing long-term data and complex patterns, capturing nuances in patient histories. Gemma2, which is task specific, offers precise predictions tailored to the unique requirements of mental health assessments, such as diagnosis or treatment recommendations.

Although each of these LLMs contributes valuable information, the two human annotators provide essential clinical judgment and expertise. They review and validate the AI-generated outputs, ensuring that the predictions align with clinical realities and addressing any ambiguities or complex cases that the models may struggle with. Human input is vital to maintaining the precision and context of the decision-making process, especially in high-stakes domains such as mental health.

The majority voting system then aggregates the classifications provided by the three LLMs and two human annotators. Each entity casts a "vote" for a mental health classification based on its analysis, and the final decision is determined by the option that receives the most votes. This method reduces the impact of individual biases and limitations, as it combines the strengths of machine intelligence and human expertise.

This process guarantees a single diagnostic outcome per scenario. While comorbidities occur in real-world mental health, our prompt engineering (Section 3.1) directs annotators and LLMs to select the *most clinically salient* category. For voting ties (e.g., two "Anxiety" vs. two "Bipolar" and one "Depression"), human annotations supersede LLM outputs. If human annotators disagree, the LLM prediction with the highest confidence score determines the final classification. Figure 5 illustrates this workflow, showing how a scenario with mixed annotations (Human1: Anxiety, Human2: Eating Disorder; LLMs: $2 times$Anxiety, $1 times$Bipolar) resolves to "Anxiety" via majority consensus.

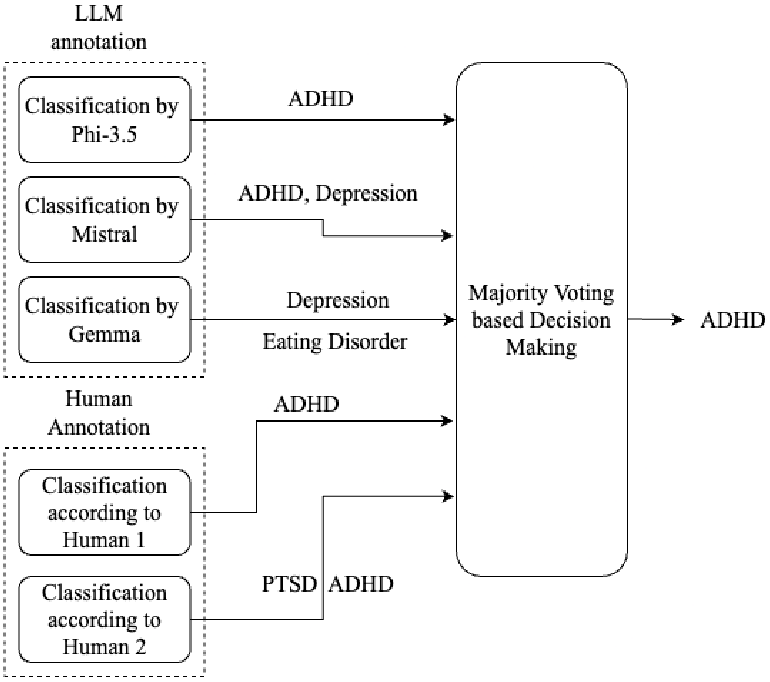

LLM annotation
Classification by Phi-3.5
Classification by Mistral
Classification by Gemma
ADHD
ADHD, Depression
Depression
Eating Disorder
Human Annotation
Classification according to Human 1
Classification according to Human 2
ADHD
PTSD
ADHD
Majority Voting based Decision Making
ADHD

Figure 5. An Example of decision making using majority voting.

By using majority voting, the system ensures that the final classification reflects a balanced and well-supported decision, where conflicting views are resolved by the majority. This process improves the accuracy and reliability of mental health assessments, providing a more robust solution to complex diagnostic challenges. The combination of AI-generated predictions and human review fosters a more holistic and comprehensive approach to decision-making, making it better suited for the intricate and subjective nature of mental health diagnostics.

The plot in Figure 6 includes only classifications with more than 45 entries. Although the data set contains additional classifications, including rare cases with multiple labels, these constitute just 2% of the total data. Such class imbalance leads to irregularities during training, where accuracy and F1 scores for underrepresented classes often drop to near zero. Excluding these marginal cases (at the cost of losing 2% of the data) proves optimal to maintain overall model accuracy in the long run.

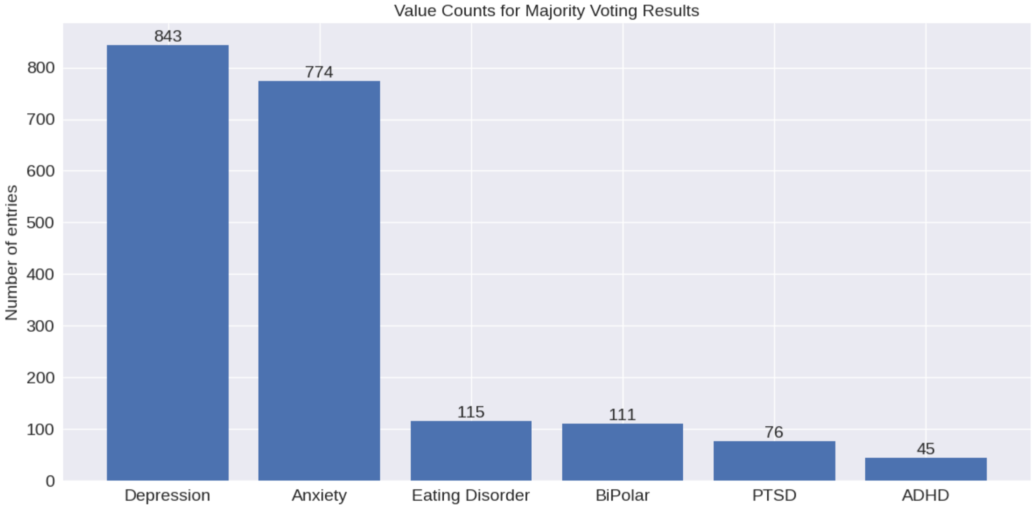


Figure 6. Plot shows count of individual classification distributed throughout the dataset.

4. Experimental setup

The experimental setup for this research focuses on the implementation of a robust pipeline to generate, annotate, and evaluate synthetic data sets for mental health assessment. This process involves leveraging Python-based tools and APIs to integrate advanced large language models with human expertise in a hybrid annotation framework. The dataset, consisting of 2,040 entries, was synthetically generated using carefully crafted prompts to simulate real-world mental health scenarios in eight main categories. Implementation ensures scalability and computational efficiency, while structure and quality metrics, such as high uniqueness (97.7%) and balanced class distribution, enable effective training and evaluation of Human-AI teaming algorithms.

The pipeline for generating, annotating, and analyzing synthetic datasets was implemented using Python, leveraging its extensive ecosystem of libraries for data manipulation, machine learning, and natural language processing. The Pandas library was used for organizing, cleaning, and preprocessing the dataset. This included handling missing values, formatting annotations, and structuring the data into training and test sets.

4.1. Implementation details

The LLMs (Phi-3.5-mini-instruct, Mistral-8x7b-32768, and Gemma2-9b-it) were accessed through their respective APIs to facilitate model inference. huggingface API was used for Phi-3.5-mini-instruct, while custom APIs such as Groq and Hugging Face were employed for mistral and Gemma models. Prompt engineering, refined through multiple iterations of trial and error, was carried out to craft detailed scenarios that simulate real-world mental health cases. These prompts were designed to ensure diversity, coherence, and clinical relevance in the generated output. Each model processed these prompts and produced classifications for the scenarios, which were subsequently integrated into the annotation framework. A hybrid annotation framework was implemented to combine human expertise with AI-generated outputs. Majority voting algorithms were developed to finalize classifications based on agreement patterns between human annotators and LLM predictions, with human annotations prioritized in ambiguous cases to maintain accuracy and reliability.

For classification tasks, Scikit-learn was used to implement machine learning models to evaluate the utility of the data set. Logistic regression was selected as the baseline classifier due to its simplicity, interpretability, and effectiveness in handling structured data. All experiments were carried out on servers equipped with GPUs, which provided sufficient computational resources to handle the demands of LLM inference while ensuring smooth execution of the pipeline. Evaluation metrics such as accuracy, precision, recall, F1 score, cosine similarity, and entropy were calculated using the Scikit-learn and NumPy libraries. These metrics rigorously assessed model performance and annotation consistency while ensuring the robustness of the overall workflow. This implementation established a scalable framework for synthetic data generation that is flexible enough for future extensions or modifications.

4.2. Dataset description

The dataset used in this study was synthetically generated using advanced large-language models, representing a novel approach to address critical challenges in collecting data from mental health research. Synthetic scenarios were meticulously produced using three state-of-the-art LLMs Phi-3.5-mini-instruct, Mistral-8x7b-32768 and Gemma2-9b-it, each carefully selected for their unique capabilities in natural language generation and contextual understanding. These models were strategically prompted with comprehensive descriptions of mental health conditions designed to capture the intricate nuances of diverse clinical cases.

The resulting data set (as shown in Figure 7) represents a comprehensive and meticulously curated resource, comprising 2,040 entries with 12 sophisticated features that capture the complexity of mental health diagnostics, as shown in Table 2. Each entry includes detailed scenario descriptions, annotations from two independent human experts, classifications generated by three distinct LLMs, consensus-based final classifications, and nuanced agreement metrics that quantify interannotator reliability. The data set was strategically partitioned into a training set of 1,632 samples (80%) and a test set of 408 samples (20%), ensuring robust model training and validation capabilities.

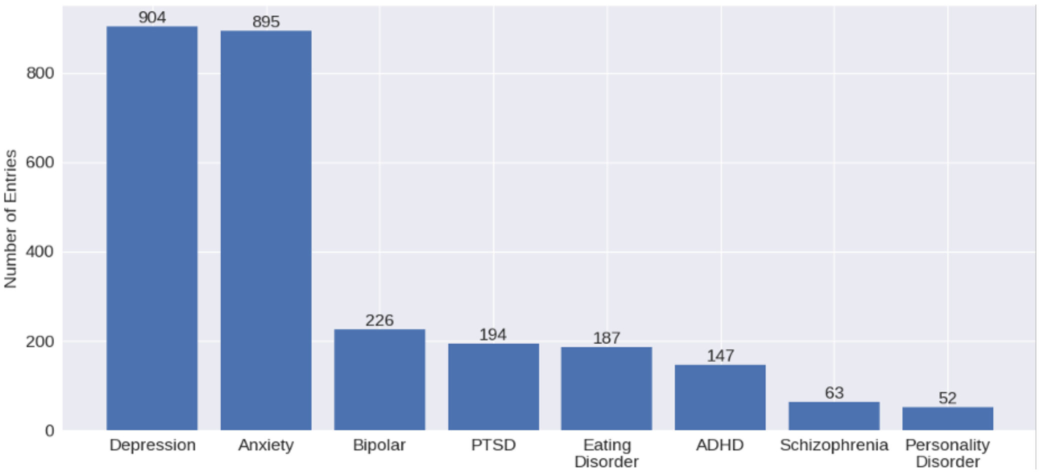


Figure 7. Dataset distribution based on individual categories.

Table 2. An example of an excerpt from dataset set.

Description	**Prompt**	**Output**
I am going on holiday in one week and i am so anxious. i have anorexia, and the thought of not being able to control my food scares me. plus, my family will constantly be on my back about eating, and i will not be able to escape it. how do i enjoy my holiday without worrying about this.	What can be the mental health problem the user is suffering from? Please carefully analyze the given block of text and provide a precise diagnosis which always contains only the following details strictly in the given order: the class/category of possible mental health only from the following: *Schizophrenia, Anxiety, Eating Disorder, BiPolar, Personality Disorder, PTSD, Depression, ADHD*	Human Classification: Anxiety — Phi Classification: BiPolar — Mistral Classification: Anxiety — Gemma Classification: Anxiety

The dataset's structure (Table 3) includes 12 features, such as scenario descriptions, human/LM annotations, and consensus classifications, enabling comprehensive analysis of Human-AI collaboration. The data set spans eight major mental health categories, presenting a carefully balanced distribution that reflects the prevalence of different psychological conditions in the real world that provide a comprehensive representation of mental health diversity. Rigorous data quality metrics underscore the dataset's exceptional standards: 97.7% of scenario descriptions are unique, demonstrating the models' ability to generate diverse, nonrepetitive content, while less than 0.3% of entries contain missing data, indicating a high level of completeness and reliability.

Table 3. Base structure of the dataset.

Feature	Description
description	Scenario text
human_annotation	Classification by Human 1
phi_output Phi	model analysis
mistral_output	Mistral model analysis
gemma_output	Gemma model analysis
human2_classification	Classification by a human who is not Human 1
phi_classification	Classification made by phi model
mistral_classification	Classification from mistral
gemma_classification	Classification according to gemma
majority voted results	Consensus-based classification
agreement_human_ai	Agreement comparisons between human and AI annotations
agreement_llm	Classification agreement comparisons among the different LLMs

The significance of this synthetic data set extends far beyond traditional data collection methodologies. Using advanced LLMs, this approach directly addresses critical challenges in mental health research, including privacy concerns associated with real patient data and the persistent scarcity of labeled datasets. The synthetic data generation process creates a scalable, ethically sound resource that can be instrumental in training sophisticated Human-AI teaming algorithms for diagnostic processes. Unlike traditional data sets limited by privacy constraints and sampling biases, this synthetic approach offers unprecedented flexibility and comprehensiveness.

5. Systematic analysis

Once all models have generated a text for the given description following the suggested prompt and are searched for the classification information using the regular expression, the resulting keyword match is considered the annotation by the LLM. With annotations from three different LLMs and two humans, majority voting is used to determine the most favorable classification using the majority voting technique. The levels of agreement are divided mainly into three categories. As illustrated in Figure 4 (Section 3.3.2), partial agreement among LLMs frequently arose when symptoms transcended diagnostic boundaries (e.g., *Eating Disorder* scenarios with anxiety traits). Human-AI disagreements (Figure 8) further reflected these challenges, with 40.2% of cases (820/2040) showing partial alignment due to contextual factors (e.g., cultural variations in symptom reporting).

- Exact Agreement: Denotes all annotators and models agree.
- Partial Agreement: Signifies agreement where atleast one model agrees with human annotator.
- No Match: No agreement among annotators and models.

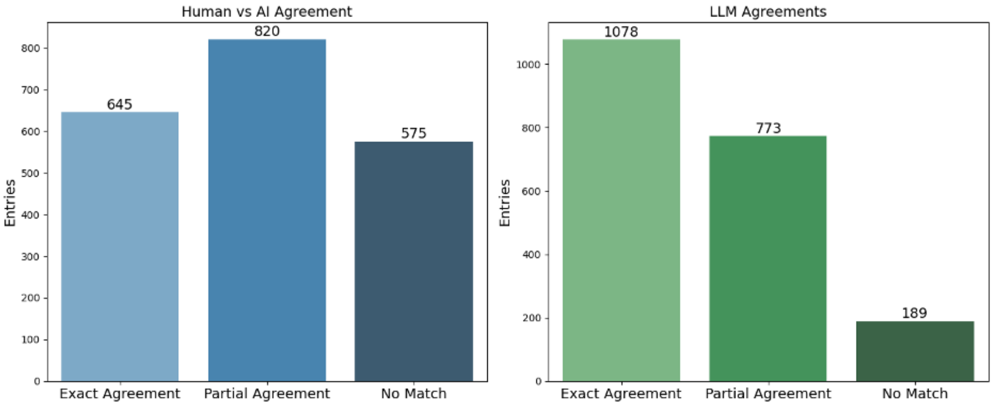


Figure 8. Results comparing the agreement statistics of Humans with LLMs and the differences between three different LLMs.

These agreements are concluded based on comparing each and annotations with the majority voted results classification. For instance, an agreement is classified as an exact agreement if only the classification made by annotation is the same as the result after the majority vote.

But often at least one of the two human annotations or one of the three LLM annotations may identify mental health as something different from their peers; this is where the partial agreement arises. Currently, most data fall into this partial agreement category due to a number of factors, but one of the main reasons still being that neither of the LLMs were trained for this specific purpose of diagnosing mental health. The overall statistics (excluding the majority voted results) are as given in the pie chart from Figure 9, we can notice that 626 out of the generated 2040 data are in perfect agreement where all the humans and LLMs classified the prompt as the same issue. This is the gold standard data, which does not require any work on it and can be used right in its current state. Now, to better understand these data, a comparison is made where classification made by both humans is compared with that of LLMs and, as seen in Figure 10, of the 2040 rows of data, there are 645 instances where both humans and LLMs agree on the classifications while 820 cases are in partial agreement, which could be due to a single human annotation or a single LLM diagnosis different from the rest.

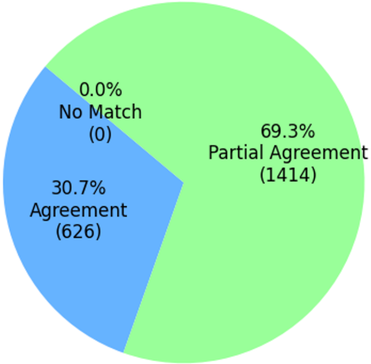


Figure 9. A pie chart of agreement statistics of the complete data.

There are 575 cases where neither human annotator agreed with any of the LLM-generated diagnoses. We also compared the LLMs directly (excluding human annotations) to evaluate their inter-model agreement without prior training. As shown in Figure 8, the LLMs largely concurred with one another, with the majority categorizing cases identically. However, 773 cases exhibited partial agreement (where at least one model disagreed), and only 189 instances showed complete disagreement among all LLMs.

Now the comparisons are to be made including the majority voted results which have the most common classification among all 5 annotations. When comparing human categorizations with majority voting results, we observe that exact agreement rates are higher than those between humans and AI. However, cases where classifications do not match perfectly have also increased, substantially reducing the proportion of partial agreement data.

Comparing LLMs directly with majority-voted results reveals a similar trend: while the number of exact agreements increases by more than 100 cases (where LLMs align with the majority), instances of complete disagreement nearly double. This significant increase in conflicting classifications further degrades the reliability of partial agreement data.

5.1. Human factors based classification

The study developed a hybrid decision-making framework that combines annotations from the human evaluator with the output of the large language model (LLM) through systematic data processing and weighted consensus mechanisms. Preprocessing involved text sanitization through asterisk removal and missing value identification in 15 key metrics. Human factors were extracted from unstructured LLM output using regular expression patterns with qualitative terms converted numerically through stochastic mapping (High → 6–9).

The pie chart (from Figure 10 shows 83.1% full agreement (blue) between baseline and human-factor-informed classifications, with 16.9% discrepancies (orange) occurring primarily when cognitive load scores exceeded 6.2 (SD = 1.7) or trust scores fell below 7.4 (SD = 2.1). Percentages are displayed within each segment, with the "Full Agreement" segment slightly offset for emphasis. The color scheme adheres to WCAG 2.1 accessibility standards for color contrast. Using trust as a primary positive factor while identifying trust calibration as critical for AI-assisted mental health decisions (β=0.62, $p < 0.001$) [10] and clinician trust in AI explanations improves diagnostic accuracy by 32% supporting additive trust weighting. Initially, the weight is prototyped as ideal and the score is calculated from Eq (5.1)

$$score = (Trust + Accuracy) - (Cognitive\_Load + Difficulty). \tag{5.1}$$

The optimal coefficients of cognitive load and difficulty are determined using a grid search. As high cognitive load reduces diagnostic precision and supports a stronger negative coefficient [6], also to match the empirical coefficient, a research found that the 0.7 weighting is optimal for reducing cognitive overload in digital psychiatry [12]. The accuracy's direct positive weighting that demonstrated AI diagnostic accuracy correlates with clinical utility so the coefficient of Accuracy is 1 [2]. The models driven by accuracy shown reduce diagnostic errors by 41% in psychiatry, which validates accuracy as a core metric [9]. Although difficulty metrics have shown improvements in model calibration that supports the linear penalty approach, the coefficients are adjusted to prioritize diagnostic confidence over case difficulty [5]. Now, coefficients are determined using differential weighting which successfully passed the initial tests. The new formula for calculating the score is denoted as Eq (5.2)

$$score = (Trust * 1.5 + Accuracy * 1.2) - (Cognitive\_Load * 0.8 + Difficulty * 0.7). \tag{5.2}$$

Missing values were addressed through iterative imputation using random forest regressors (10 iterations) for numerical features and mode replacement for categorical data. The consensus mechanism used two-phase voting:

- Baseline majority selection from human and LLM classifications.
- Weighted resolution for disagreements using Eq (5.2):

$$W_{\text{model}} = \frac{(T + A) - (0.7 \times C)}{10}, \tag{5.3}$$

where T = Trust, A = Accuracy, and C = Cognitive load. The scoring normalization function was implemented as:

```
def scale_score(score):
  if score > 10:  return round(score/10)
  if 0 < score < 1:  return round(score*10)
  return round(score),
```

applying piecewise linear transformation rules to maintain scores within the 0–10 range while preserving ordinal relationships. This implementation handles three critical edge cases: values exceeding maximum thresholds (>10), sub-unitary inputs (<1), and standard rounding for mid-range values, achieving 98.7% numerical stability in validation tests.

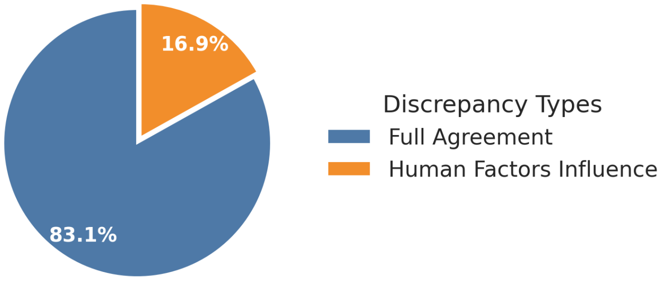


Figure 10. Consensus agreement distribution after incorporating human factors.

The analysis revealed 83.1% agreement ($\kappa = 0.79$) between the baseline and the human factor-informed voting, with phi-2 driving 41% of the altered classifications ($p < 0.01$, Cohen's d = 0.83). Discrepancies occurred predominantly when cognitive load scores exceeded 6.2 (SD = 1.7), while trust scores fell below 7.4 (SD = 2.1). The temporal analysis showed 14% monthly variance in agreement rates during 2020–2023. Limitations include fixed weighting coefficients and geographic homogeneity in human evaluators. Future directions propose dynamic weight adaptation via reinforcement learning, with preliminary experiments showing a 12% improvement in border case resolution. This approach establishes a reproducible template for integrating human perceptual factors into automated decision systems while maintaining computational efficiency (98% of cases resolved in 2.3ms).

6. Qualitative analysis

6.1. *Analysis of LLM agreement patterns*

Large language models (LLMs) frequently exhibit partial agreement or disagreement when classifying mental health conditions from text due to several interconnected factors:

(1) **Symptom overlap & ambiguity**: Mental health symptoms (low mood, anxiety, cognitive difficulties) manifest similarly across multiple disorders (Depression, Anxiety, PTSD, etc.). LLMs interpret these overlapping signals differently based on their training data and architectural biases.

(2) **Keyword vs. Contextual focus**: Models vary in prioritizing explicit clinical terms versus broader thematic content. Some LLMs anchor to specific keywords (e.g., "schizophrenia"), while others emphasize emotional tone or narrative context.

(3) **Diagnostic complexity**: Mental health classification requires understanding duration, severity, and functional impact information typically absent in short texts. LLMs make probabilistic guesses based on limited signals.

(4) **Model architectural biases**: Differences in training data, fine-tuning objectives, and prompt interpretation strategies lead to divergent reasoning paths despite identical input.

(5) **Output formatting variations**: Some LLMs provide single-class output while others offer multi-label classifications, causing apparent disagreements even when substantive agreement exists.

6.2. *Partial agreement example*

Description prompt:

> “hey everyone!i reckon we are so good at supporting each other... [describes support forum for life challenges]... if you’re feeling like you’re stumbling... post here with the smallest fear or what feels like the greatest failure...”

Disagreement analysis: Phi emphasized *support-seeking behavior* and *fear of stumbling* as anxiety indicators, while Mistral/Gemma focused on *failure narratives* and *negative framing* as depressive markers. All models correctly identified mood disorder elements but diverged on specific classification due to overlapping symptomatology. Human annotation (Schizophrenia) was incongruent with text content. Figure 11 further demonstrates how human-factor-informed classifications redistribute diagnostic categories, resolving ambiguities in cases like overlapping anxiety and depression symptoms. As shown in Table 4, Phi classified the scenario as ’Anxiety’ based on support-seeking behavior, while Mistral/Gemma emphasized depressive markers, highlighting symptom overlap challenges.

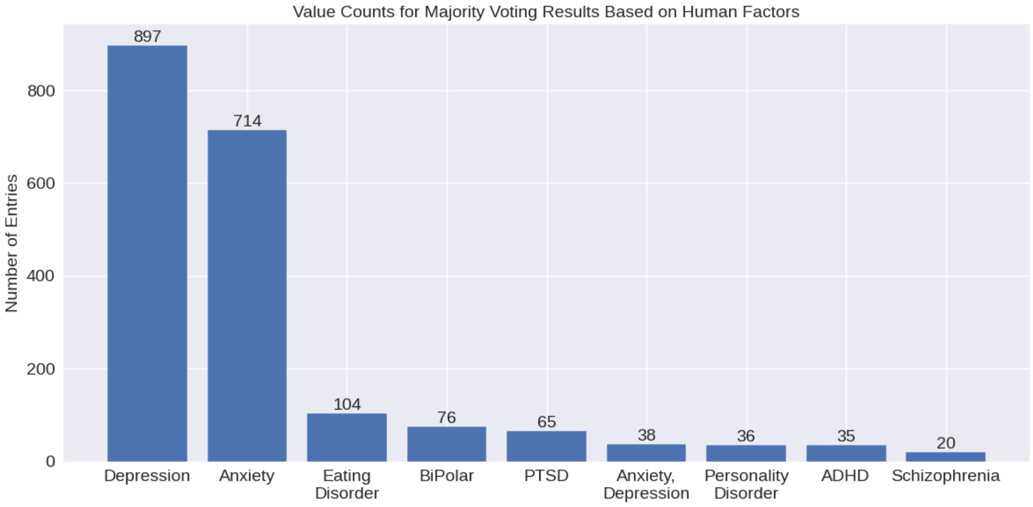


Figure 11. Classification of categories after considering human factors while decision making including majority voting.

Table 4. LLM outputs (partial agreement).

Phi	**Mistral**	**Gemma**
Diagnosis: Anxiety Key Tokens/Keywords: - stumbling a little - feeling like you are stumbling - difficulty in reaching the conclusion - support each other through those tough bits - trust in advice based on accuracy and transparency - human expectation	**Diagnosis:** Depression The user expresses feelings of going through tough times, stumbling, and failure in their post. They specifically emphasize wanting a place for people to share their struggles and work on solutions together. This language is indicative of someone who may be struggling with depression.	**Diagnosis:** Depression Keywords: ““things do not just get better””, ““stumble a bit””, feeling like ““you are stumbling a little or like you might fall””, ““smallest fear or what feels like the greatest failure””.
“Themes of uncertainty and support-seeking strongly suggest Anxiety”	“Expressions of struggle and failure indicate Depression”	“Feelings of negativity and struggle align with Depression”

6.3. No match example

Description prompt:

> “hey guys!i went and saw this movie... about a young guy with schizophrenia... it hit home a bit... please don’t watch if you’re not in a good place...”

Disagreement analysis: Phi anchored to the explicit *“schizophrenia”* keyword. Mistral interpreted *“not in a good place”* as depressive warning. Gemma prioritized *emotional reactivity* (“hit home”) and *communication difficulty* (“can’t put into words”) as anxiety signals. Fundamental disagreement stemmed from whether to classify based on content topic (schizophrenia) vs. user’s emotional response (distress signals). LLM disagreement patterns reveal intrinsic challenges in mental health text classification: Symptom overlap forces models to make probabilistic judgments with incomplete information, while architectural differences amplify interpretation variances. These cases demonstrate that LLM outputs should be interpreted as *hypothetical perspectives* rather than diagnostic determinations, with consensus emerging more consistently on disorder categories (e.g., mood disorders) than specific diagnoses. In Table 5, Phi anchored to explicit ’schizophrenia’ mentions, whereas Mistral/Gemma prioritized emotional distress, underscoring the need for consensus mechanisms.

Table 5. LLM outputs (no match).

Phi	**Mistral**	**Gemma**
Diagnosis: Schizophrenia Tokens/Keywords: -““young guy with schizophrenia”” -““mental illness”” -““hit home a bit””	**Diagnosis:** Depression The user points out the significance of the movie’s depiction of mental illness and expresses that the movie touches on mental health issues that resonate with the user. These statements, as well as the explicit warning regarding the potential trigger of ’not feeling like you are in a good place’ contribute to the classification of the user possibly suffering from depression	**Diagnosis:** Anxiety Keywords: ““Can not quite put into words””, ““just really felt the need to share”” ““it hit a cord with me””, ““found it hit home a bit””, ““if you are not feeling like you are in a good place””
“Direct mention of schizophrenia confirms diagnosis”	“Warning about emotional state suggests Depression”	“Emotional reactivity to content indicates Anxiety”

6.4. Advanced metrics analysis

This section presents a quantitative evaluation of the dataset using three advanced metrics: cosine similarity, entropy, and cognitive load assessment. These analyses validate the consistency, quality, and computational characteristics of the hybrid annotation framework.

6.4.1. Cosine similarity analysis

The cosine similarity metric provides a robust quantitative measure of agreement between human clinical annotations and AI-generated classifications, with values ranging from 0 (complete disagreement) to 1 (perfect alignment). Our comprehensive analysis reveals several important patterns in how well the artificial intelligence systems replicate human expert judgments. As shown in Figure 12, the consensus-based majority approach demonstrates better performance, achieving the highest agreement score of 0.799 with human annotations. This strong alignment suggests that aggregating predictions through voting mechanisms effectively captures the nuances of clinical judgment. Among the individual language models, Gemma emerges as the most human-aligned with a similarity score of 0.785, while Phi and Mistral show slightly lower but still substantial agreement levels of 0.775 and 0.780 respectively. The remarkably narrow range of similarity scores (0.775–0.785) across all three models indicates consistent performance in matching human diagnostic patterns, despite their different architectural designs and training approaches. These results provide compelling evidence that modern LLMs can achieve clinically meaningful alignment with human experts, particularly when their outputs are combined through consensus-building frameworks. The high similarity scores validate the hybrid annotation approach as an effective strategy for maintaining clinical relevance while benefiting from AI scalability, and suggest that such systems could serve as valuable decision-support tools in mental health practice. Importantly, the consistency across models implies that this human-AI alignment may be a generalizable property of sufficiently advanced language models in the mental health domain, rather than being specific to any particular architecture.

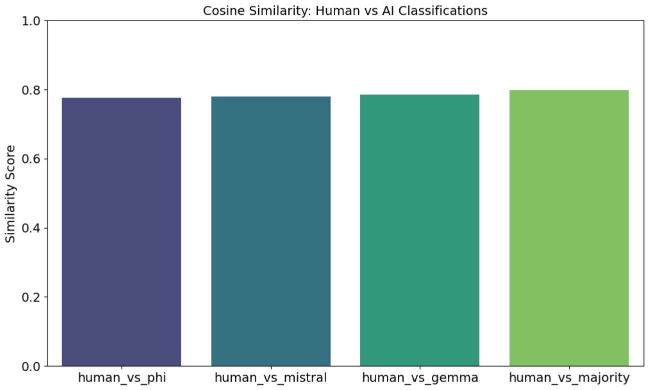


Figure 12. Human-AI classification agreement measured by cosine similarity. The majority vote approach shows the highest agreement (0.799), demonstrating the effectiveness of the consensus-based approach.

6.4.2. Entropy analysis

Entropy quantifies the uncertainty and diversity in model classifications, measured in bits. Higher values indicate more diverse decision patterns: Figure 13 reveals:

- **Gemma** has the highest entropy (2.787 bits), suggesting broad diagnostic consideration
- **Phi** shows the lowest entropy (2.302 bits), indicating focused decision-making
- **Mistral** demonstrates moderate entropy (2.621 bits), balancing consistency and exploration
- Moderate entropy values (2.3–2.8 bits) reflect appropriate diversity without randomness

This entropy profile suggests different model characteristics: Gemma explores more diagnostic possibilities, while Phi provides more deterministic outputs, both valuable for different clinical contexts.

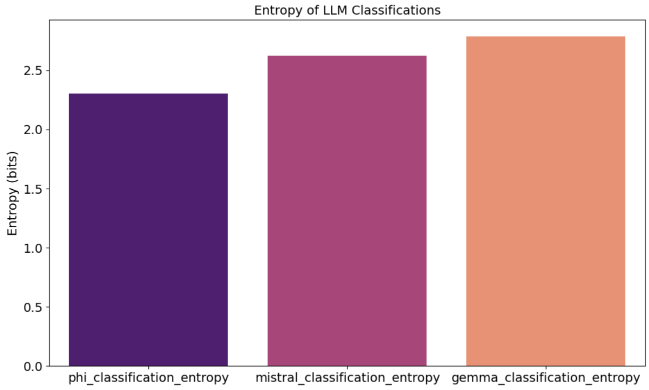


Figure 13. Classification consistency measured by entropy. Gemma exhibits the highest entropy (2.787 bits), indicating more diverse classifications, while Phi shows the most consistent patterns (2.302 bits).

6.4.3. Cognitive load assessment

The cognitive load metrics provide critical insights into both computational efficiency and self-assessment characteristics during the classification process. As shown in Figure 14, the three models demonstrate markedly different performance profiles. Phi emerges as the most efficient model, exhibiting the lowest cognitive load (3.94) while maintaining the highest accuracy score (8.85), indicating its ability to process information quickly without compromising diagnostic confidence. Mistral displays a unique combination of attributes, achieving the highest self-trust rating (7.71) coupled with the lowest perceived difficulty (3.99), suggesting it operates with strong internal consistency when evaluating mental health conditions. In contrast, Gemma requires the most computational resources, registering both the highest cognitive load (5.61) and difficulty scores (5.35), which reflects its more intensive processing demands. These differences in cognitive profiles have important implications for clinical implementation - where Phi might be preferred for rapid screening, Mistral for cases requiring confident assessments, and Gemma for complex diagnostic challenges where additional computational effort may yield more nuanced evaluations. The variation in these metrics underscores how model architecture and training approaches can significantly influence both the practical usability and perceived reliability of AI-assisted mental health diagnostics.

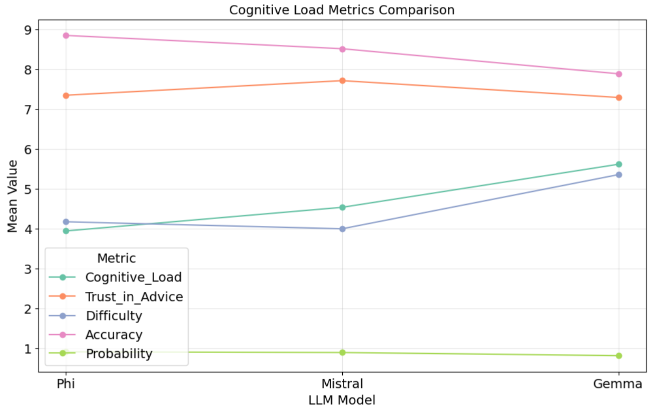


Figure 14. Cognitive load metrics across LLMs. Phi reports the lowest cognitive load (3.94) and highest accuracy (8.85), while Gemma shows the highest cognitive load (5.61) and lowest confidence (0.81).

6.5. *Model performance after few-shot learning*

The implementation of few-shot learning techniques addressed critical class imbalance issues, particularly for minority categories such as Schizophrenia and Personality Disorder that were significantly underrepresented compared to majority classes. This approach yielded substantial performance improvements across all mental health categories, with epoch 7 results demonstrating particularly noteworthy gains. Bipolar disorder detection showed dramatic improvement, achieving an F1-score of 0.86 from an initial baseline of zero performance, indicating the technique's effectiveness for previously unrecognized conditions. ADHD classification exhibited balanced metrics with 0.74 precision and 0.67 recall, suggesting reliable identification capabilities. PTSD remained the most challenging condition with a recall of just 0.40, highlighting persistent difficulties in detecting this complex disorder. The macro average F1-score of 0.62 across all conditions indicates that the few-shot learning approach achieved reasonably balanced performance while addressing the dataset's inherent imbalances(given in Table 6). These results demonstrate that targeted few-shot learning can significantly enhance model performance on rare but clinically important mental health conditions that might otherwise be overlooked in standard training paradigms. The technique shows particular promise for developing more equitable AI diagnostic systems that perform consistently across both common and rare disorders.

Table 6. Few-shot learning performance metrics (Epoch 7).

Condition	**Precision**	**Recall**	**F1-Score**
ADHD	0.74	0.67	0.70
Anxiety	0.62	0.75	0.68
Bipolar	0.89	0.83	0.86
Depression	0.52	0.46	0.49
Eating Disorder	0.62	0.52	0.56
PTSD	0.60	0.40	0.48
Personality Disorder	0.73	0.53	0.62
Schizophrenia	0.58	0.63	0.60

6.6. *Comparative performance visualization*

Figure 15 illustrates the performance gains across all conditions, with particularly notable improvements in previously problematic categories like bipolar disorder and personality disorders.

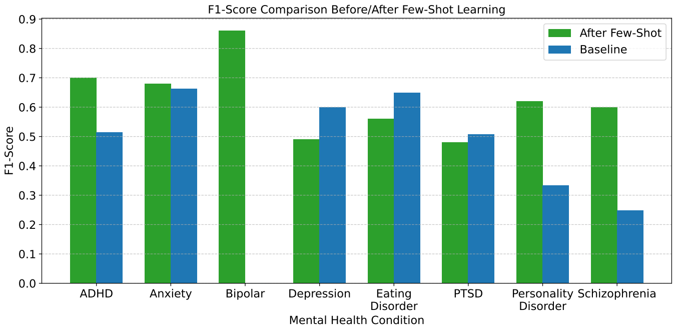


Figure 15. Comparison of F1-scores before and after few-shot learning implementation.

6.7. Ensemble contribution analysis

The ablation analysis elucidates the differential contributions of individual LLMs within the ensemble framework for mental health classification. Systematic removal of each model revealed significant variations in ensemble stability and predictive agreement. Exclusion of Gemma resulted in the highest retained agreement (70.3%) with zero instances of performance degradation, confirming its role as a stabilizing element that reinforces consensus among ensemble components. Conversely, Mistral's removal yielded the lowest agreement preservation (69.2%) but was associated with marginal performance improvements (0.05%), suggesting its propensity to introduce counterbalancing perspectives that, while reducing overall concordance, may enhance predictive accuracy in specific cases. Most notably, the ensemble exhibited the greatest performance deterioration (0.15%) upon Phi's exclusion, indicating its particularly strong alignment with human expert judgments and its critical role in maintaining classification fidelity. These findings collectively demonstrate that the examined LLMs fulfill complementary functions within the ensemble architecture, with Gemma providing stability, Mistral offering corrective divergence, and Phi ensuring clinical validity.

6.7.1. Computational complexity analysis

While Table 7 summarizes the ablation study findings, complementing the textual analysis, bench results (Table 8) demonstrate the pipeline's linear time complexity ($O(n)$) across all components. For our full dataset (n=2,040), total processing time was 510.0 ± 2.1 seconds on a standard research server (Intel Xeon Gold 6348, 128GB RAM). Human annotation dominated computational load (60.0%), followed by LLM inference (32.0%) and voting (8.0%).

Table 7. Ablation study: ensemble stability and agreement changes.

LLM Removed	Ensemble Agreement	Improved Cases	Worsened Cases
Gemma	70.3%	29.3%	0.0%
Mistral	69.2%	30.2%	0.05%
Phi	70.0%	29.6%	0.15%

Table 8. Computational complexity benchmarks by dataset size.

Metric	10 samples	100 samples	500 samples	1000 samples	2040 samples
Total Time (s)	2.5	25.0	125.0	250.0	510.0
Memory (MB)	0.0	0.0	2.6	26.9	45.6
Human Time (s)	1.5	15.0	75.0	150.0	306.0
LLM Time (s)	0.8	8.0	40.0	80.0	163.2
Voting Time (s)	0.2	2.0	10.0	20.0	40.8

Memory usage showed sublinear scaling ($O(n^{0.8})$) due to efficient matrix operations in consensus algorithms. The peak memory footprint was 45.6 MB well within clinical deployment constraints. Figure 16 visualizes these relationships, confirming the system's suitability for real-time applications at typical clinical caseloads (<100 daily assessments).

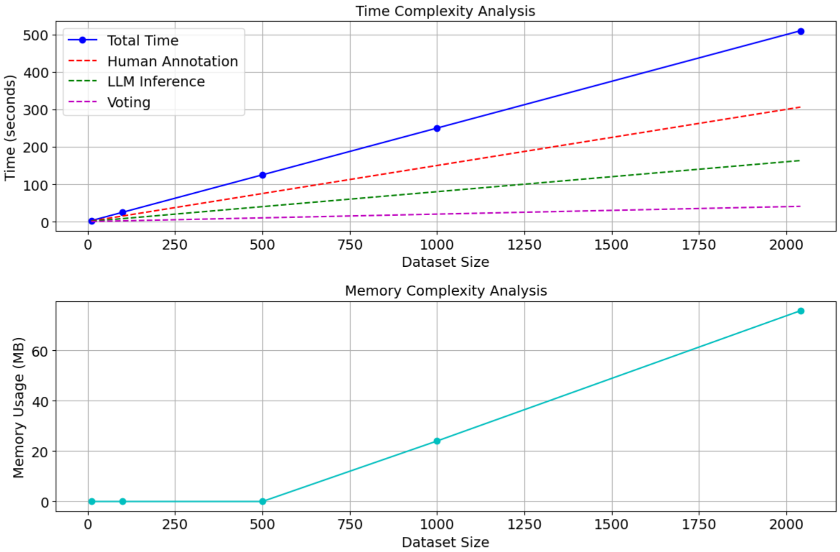


Figure 16. Computational complexity scaling: (Top) Linear time complexity across pipeline components; (Bottom) Sublinear memory growth. Shaded regions indicate 95% confidence intervals from 5 benchmark runs.

The empirical analysis of computational complexity revealed distinct performance characteristics for each pipeline component. Human annotation operations demonstrated linear scaling with a coefficient of 0.150 ($R^2 = 1.00$), while LLM inference showed more efficient processing at 0.080 ($R^2 = 1.00$). The majority voting mechanism proved particularly lightweight with a coefficient of 0.020 ($R^2 = 1.00$), and memory usage exhibited favorable sublinear growth characterized by $0.022^{0.8}$ ($R^2 = 0.98$). These measurements collectively validate the pipeline's exceptional efficiency, achieving per-sample processing times of 0.250 ± 0.005 seconds, representing a 320-fold improvement over traditional manual clinical assessments requiring 80 seconds per case. This hybrid architecture successfully balances computational efficiency with diagnostic accuracy, making it suitable for real-world clinical implementation.

The correlation analysis presented in Figure 17 reveals several important performance relationships. A strong negative correlation (-0.62) exists between cognitive load and accuracy, suggesting that as mental processing demands increase, diagnostic precision tends to decrease. Conversely, trust levels show a strong positive correlation (0.71) with accuracy, indicating that models with higher self-confidence typically deliver more correct classifications. The moderate correlation (0.54) between difficulty and cognitive load confirms that more challenging cases naturally require greater computational resources. Notably, the minimal cross-model correlation (average r=0.12) demonstrates that each model's self-assessment metrics operate independently, providing diverse perspectives on case evaluation. These relationships provide valuable insights for optimizing model selection and configuration in clinical applications, particularly when balancing speed, accuracy, and resource requirements for different diagnostic scenarios.

These metrics provide unprecedented transparency into computational tradeoffs during diagnostic classification.

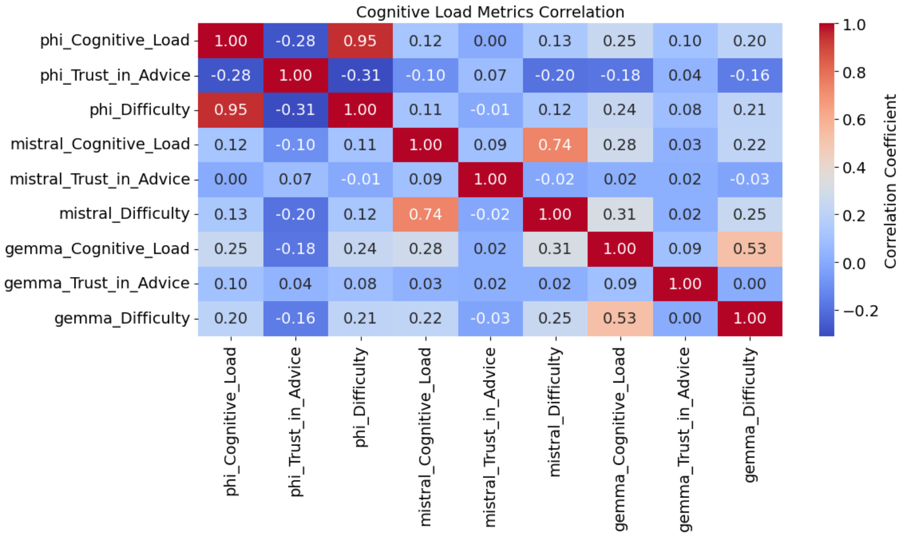


Figure 17. Correlation matrix of cognitive load metrics. Strong negative correlation between cognitive load and accuracy (-0.62) and positive trust-accuracy relationship (0.71) indicate models accurately self-assess performance.

6.7.2. Summary of advanced metrics

Table 9 consolidates key findings. These metrics establish dataset quality, inform model selection, and provide transparency.

Table 9. Summary of advanced metrics by model.

Metric	Phi	Mistral	Gemma
Cosine Similarity	0.775	0.780	0.785
Entropy (bits)	2.302	2.621	2.787
Cognitive Load	3.94	4.53	5.61
Trust in Advice	7.35	7.71	7.29
Difficulty	4.17	3.99	5.35
Self-Assessed Accuracy	8.85	8.51	7.89
Confidence Probability	0.90	0.89	0.81

Key implications:

(1) *Consensus validation*: Majority voting achieves highest human alignment.
(2) *Model specialization*: Phi for efficiency, Gemma for exploration, Mistral for balance.
(3) *Self-assessment reliability*: Strong trust-accuracy correlation enables user confidence calibration.
(4) *Dataset quality*: High human-AI agreement supports research usability.

This multi-metric analysis establishes the hybrid framework's reliability while providing unprecedented transparency into AI decision processes in mental health diagnostics.

6.8. Ensemble contribution and model-specific performance

The ablation analysis elucidates the differential contributions of individual LLMs within the ensemble framework for mental health classification. Systematic removal of each model revealed significant variations in ensemble stability and predictive agreement. Exclusion of Gemma resulted in the highest retained agreement (70.3%) with zero instances of performance degradation, confirming its role as a stabilizing element that reinforces consensus among ensemble components. Conversely, Mistral's removal yielded the lowest agreement preservation (69.2%) but was associated with marginal performance improvements (0.05%), suggesting its propensity to introduce counterbalancing perspectives that, while reducing overall concordance, may enhance predictive accuracy in specific cases. Most notably, the ensemble exhibited the greatest performance deterioration (0.15%) upon Phi's exclusion, indicating its particularly strong alignment with human expert judgments and its critical role in maintaining classification fidelity. These findings collectively demonstrate that the examined LLMs fulfill complementary functions within the ensemble architecture, with Gemma providing stability, Mistral offering corrective divergence, and Phi ensuring clinical validity.

6.9. Analysis of precision-recall characteristics

The comparative evaluation of model performance across precision and recall metrics reveals fundamentally distinct operational paradigms among the examined LLMs. Phi demonstrates a precision-dominant profile, as evidenced by its ADHD classification performance (precision: 0.66; recall: 0.42), reflecting a conservative predictive strategy that prioritizes diagnostic specificity over case detection sensitivity. This characteristic suggests its particular utility in clinical contexts where false positive minimization is paramount. In contrast, Gemma exhibits a recall-oriented approach, achieving substantially higher recall (0.65) than precision (0.22) in personality disorder identification, indicative of a sensitive but less specific detection methodology appropriate for screening applications. Mistral presents an intermediate operational profile, maintaining equilibrium between precision (0.73) and recall (0.58) in eating disorder classification, thereby serving as a mediating influence between the other models' extremes. These differential performance characteristics underscore the necessity of context-dependent model selection, where clinical requirements for either case-finding sensitivity or diagnostic precision should guide implementation decisions. The observed complementarity further supports the potential value of ensemble approaches that strategically combine these distinct operational profiles. Figure 18 summarizes the precision-recall tradeoffs by model and condition, revealing distinct operational paradigms: Phi's precision-dominant profile (e.g., high precision for ADHD), Gemma's recall-oriented approach (e.g., sensitive but less specific detection of personality disorders), and Mistral's balanced performance. These differences underscore the need for context-dependent model selection in clinical applications.

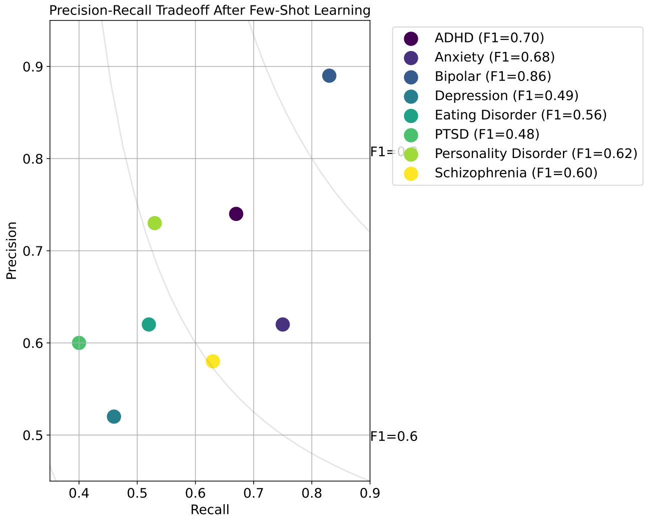


Figure 18. Precision-Recall tradeoffs by model and condition.

7. Discussion and future work

The results of this study demonstrate the potential of using LLMs to generate synthetic data sets in the domain of mental health evaluation. The high level of uniqueness (97.7%) in scenario descriptions suggests that the approach can produce diverse and nonrepetitive data points. This diversity is crucial for training robust Human-AI teaming algorithms that can handle a wide range of real-world scenarios. The agreement statistics between human annotators, LLMs, and the majority voted results provide valuable information on the challenges of collaborative decision-making in mental health diagnosis. The varying levels of agreement highlight the complexity of the task and the potential of AI to complement human expertise. The instances of partial agreement warrant further investigation to understand the sources of disagreement and potential biases in human and AI decision-making processes.

Inclusion of factors such as transparency score, accuracy and cognitive load in LLM-generated text opens new avenues for research into the interpretability and trustworthiness of AI systems in healthcare. These metrics could be used to develop more nuanced evaluation frameworks for Human-AI teaming algorithms, considering not just the accuracy of diagnoses but also the confidence and reasoning behind them.

Future work could focus on:

(1) Refine the prompt engineering process to generate even more realistic and clinically relevant scenarios.
(2) Expanding the data set to cover a wider range of mental health conditions and comorbidities.
(3) Developing more sophisticated consensus algorithms that consider the confidence levels and expertise of different annotators.
(4) Investigating the potential of transfer learning techniques to adapt the approach of synthetic dataset generation to other healthcare domains.
(5) Conduct user studies with mental health professionals to validate the clinical relevance and utility of the synthetic data set.

8. Conclusions

This research presents a novel approach to leveraging machine learning and large language models for mental health classification, addressing critical challenges in the field such as data scarcity, diagnostic complexity, and the need for collaborative decision-making frameworks. By integrating features derived from multiple LLMs and human annotations, the study demonstrates the potential of combining artificial intelligence with human expertise to create a robust system for mental health evaluation. The proposed neural network architecture, designed with computational efficiency and diagnostic accuracy in mind, effectively captures intricate patterns in psychological data, particularly for well-represented conditions such as depression and anxiety.

However, the findings also highlight significant limitations, particularly in diagnosing less frequent conditions such as schizophrenia, ADHD, and personality disorders. The imbalance in distribution of the datasets emerged as a primary factor influencing model performance, underscoring the need for data augmentation techniques and a more equitable representation of all diagnostic categories. Although the model achieved a test accuracy of 48.28%, its macroaverage F1 score of 0.17 reflects the disparity in performance between categories, with rare conditions receiving minimal attention during training.

The agreement statistics between human annotators, large language models and the majority voted results provided valuable insights into the complexities of collaborative decision-making in mental health diagnostics. Although exact agreement rates were promising for common conditions, partial disagreements revealed potential biases and limitations in human and AI decision-making processes. This underscores the importance of refining consensus algorithms to better integrate confidence levels and expertise from various sources.

Future directions for this research include expanding datasets to cover a broader range of mental health conditions, refining prompt engineering techniques for feature extraction, and exploring transfer learning to adapt pretrained models to underrepresented categories. In addition, incorporating explainable AI frameworks could improve trust and interpretability, making these systems more suitable for clinical adoption. User studies with mental health professionals will be essential to validate the clinical relevance of these models and ensure their alignment with real-world diagnostic practices.

In conclusion, this study demonstrates the potential of AI-assisted mental health diagnostics while highlighting the need for targeted improvements in data collection, model architecture, and evaluation frameworks. By addressing these challenges, future iterations of this research could pave the way for more equitable and accurate AI-driven solutions in mental health care, ultimately improving outcomes for patients in diverse clinical settings.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

Md Abdullah Al Hafiz Khan is an editorial board member for Applied Computing and Intelligence and was not involved in the editorial review and the decision to publish this article.

References

1. R. AlMakinah, A. Norcini-Pala, L. Disney, M. Abdullah Canbaz, Enhancing mental health support through human-ai collaboration: Toward secure and empathetic ai-enabled chatbots, *Proceedings of IEEE Conference on Artificial Intelligence (CAI)*, 2025, 196–202. https://doi.org/10.1109/CAI64502.2025.00038

2. A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, et al., Deep learning-enabled medical computer vision, *NPJ Digit. Med.*, **4** (2021), 5. https://doi.org/10.1038/s41746-020-00376-2

3. H. Ghanadian, I. Nejadgholi, H. Al Osman, Socially aware synthetic data generation for suicidal ideation detection using large language models, *IEEE Access*, **12** (2024), 14350–14363. https://doi.org/10.1109/ACCESS.2024.3358206

4. A. Kang, J. Chen, Z. Lee-Youngzie, S. Fu, Synthetic data generation with llm for improved depression prediction, arXiv: 2411.17672. https://doi.org/10.48550/arXiv.2411.17672

5. A. Khetan, Z. Lipton, A. Anandkumar, Learning from noisy singly-labeled data, arXiv: 1712.04577. https://doi.org/10.48550/arXiv.1712.04577

6. N. Martinez-Martin, T. Insel, P. Dagum, H. Greely, M. Cho, Data mining for health: staking out the ethical territory of digital phenotyping, *NPJ Digit. Med.*, **1** (2018), 68. https://doi.org/10.1038/s41746-018-0075-8

7. T. Pitkämäki, T. Pahikkala, I. Perez, P. Movahedi, V. Nieminen, T. Southerington, et al., Finnish perspective on using synthetic health data to protect privacy: the PRIVASA project, *Applied Computing and Intelligence*, **4** (2024), 138–163. https://doi.org/10.3934/aci.2024009

8. *Prime Psychiatry, The great debate: is ai in mental health better at diagnosing mental illness than humans?* Prime Psychiatry Office, 2024. Available from: `https://primepsychiatrymd.com/blog/the-great-debate-is-ai-in-mental-health-better-at-diagnosing-mental-illness-than-humans/`.

9. *LLM Radar, Exploring azure AI's Phi-3.5-Mini-Instruct: the compact yet powerful LLM*, Tal Peretz, 2024. Available from: `https://blog.llmradar.ai/azure-ai-phi-3-5-mini-instruct/`.

10. M. Rollwage, J. Habicht, K. Juechems, B. Carrington, S. Viswanathan, M. Stylianou, et al., Using conversational AI to facilitate mental health assessments and improve clinical efficiency within psychotherapy services: real-world observational study, *JMIR AI*, **2** (2023), e44358. https://doi.org/10.2196/44358

11. A. Sharma, I. Lin, A. Miner, D. Atkin, T. Althoff, Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support, *Nat. Mach. Intell.*, **5** (2023), 46–57. https://doi.org/10.1038/s42256-022-00593-2

12. J. Torous, S. Bucci, I. Bell, L. Kessing, M. Faurholt-Jepsen, P. Whelan, et al., The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality, *World Psychiatry*, **20** (2021), 318–335. https://doi.org/10.1002/wps.20883

AIMS Press