



NEURAL INFORMATION
PROCESSING SYSTEMS



武汉大学
WUHAN UNIVERSITY

MoodAngels: A Retrieval-augmented Multi-agent Framework for Psychiatry Diagnosis

Mengxi Xiao, Ben Liu, He Li, Jimin Huang, Qianqian Xie, Xiaofen Zong*, Mang Ye*, Min Peng*

Presenter: Mengxi Xiao

2025.10.24



NEURAL INFORMATION
PROCESSING SYSTEMS

Research Background



武漢大學

WUHAN UNIVERSITY

Three Core Challenges of AI in Psychiatric Diagnosis

Strong Subjectivity in Assessment

Psychiatric diagnosis relies on self-reports and clinical evaluations, which are prone to individual cognitive biases and lack objective biomarkers.

High Symptom Overlap

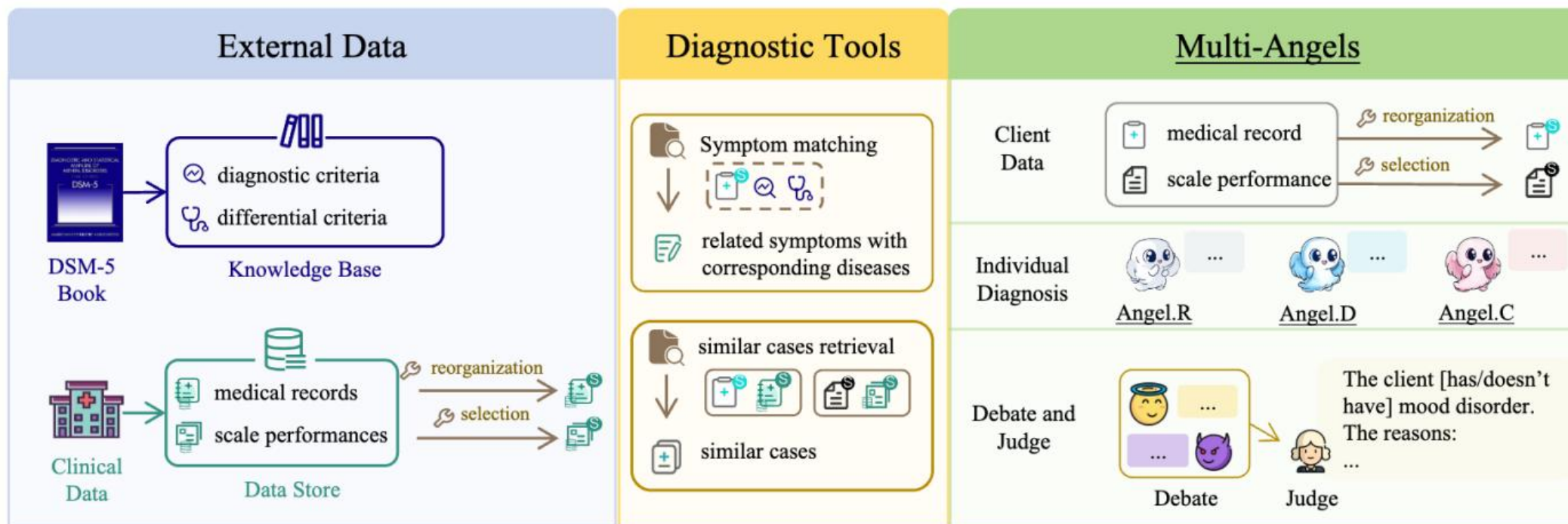
Mood disorders (depression, bipolar disorder) share overlapping symptoms with other psychiatric conditions (e.g., schizophrenia, anxiety disorders), leading to misdiagnosis.

Data Privacy Bottleneck

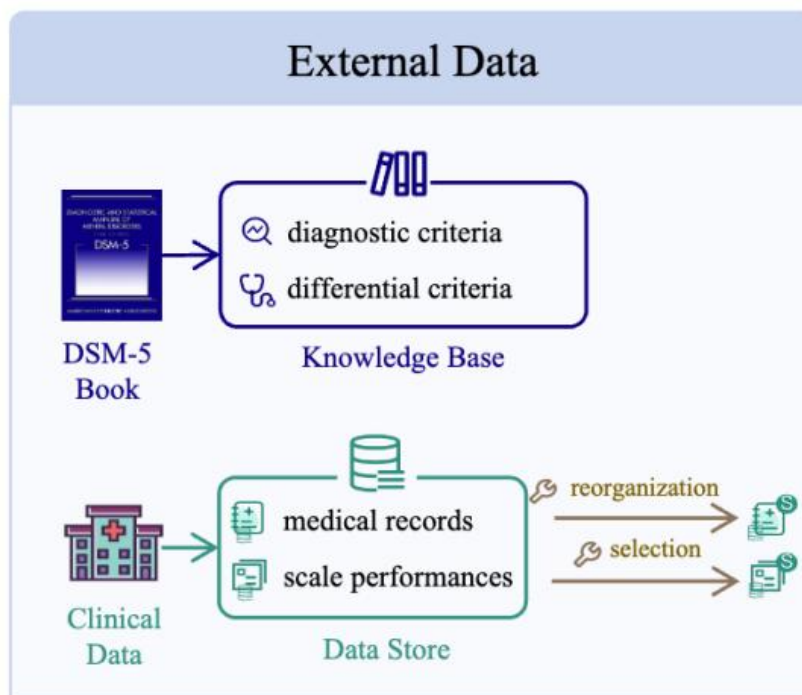
Real clinical data contains sensitive information and cannot be publicly shared, resulting in scarcity of training data for AI models.

MoodAngels: A Retrieval-augmented Multi-agent Framework for Psychiatry Diagnosis

Framework Design: Three-Layer Core Mechanism



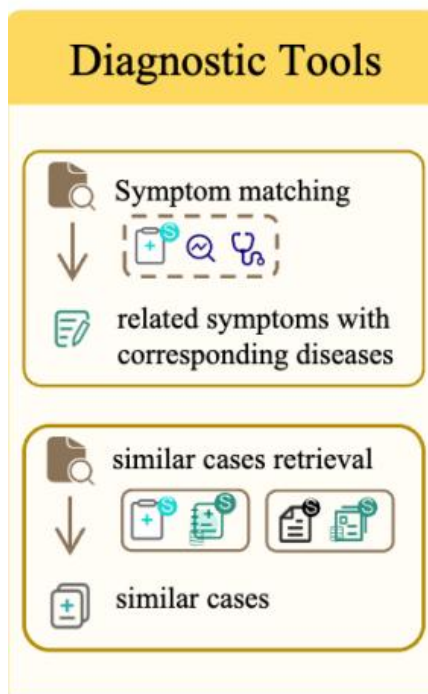
Framework Design: Three-Layer Core Mechanism



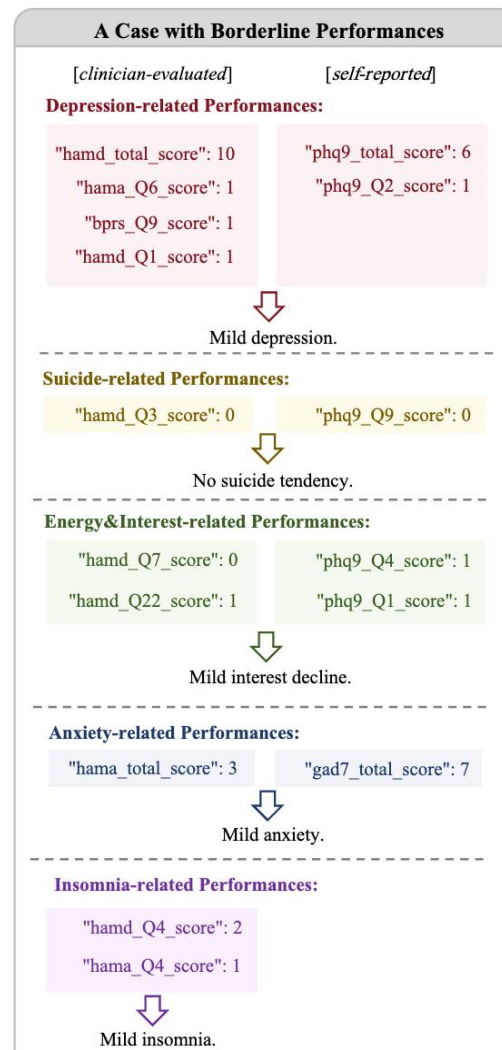
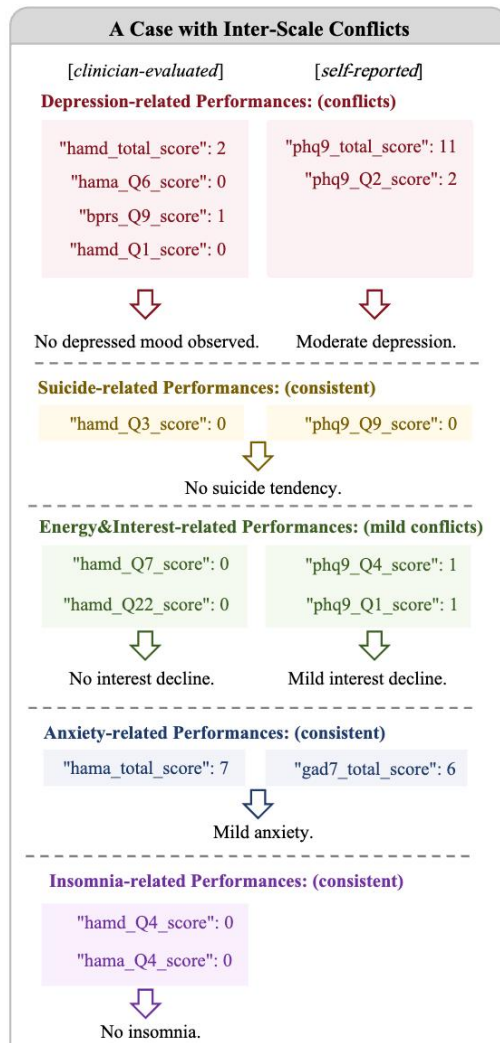
Knowledge Base

Structured DSM-5 diagnostic criteria + 2,804 anonymized real clinical cases, supporting symptom matching and similar case retrieval (using BGE-M3 semantic embedding).

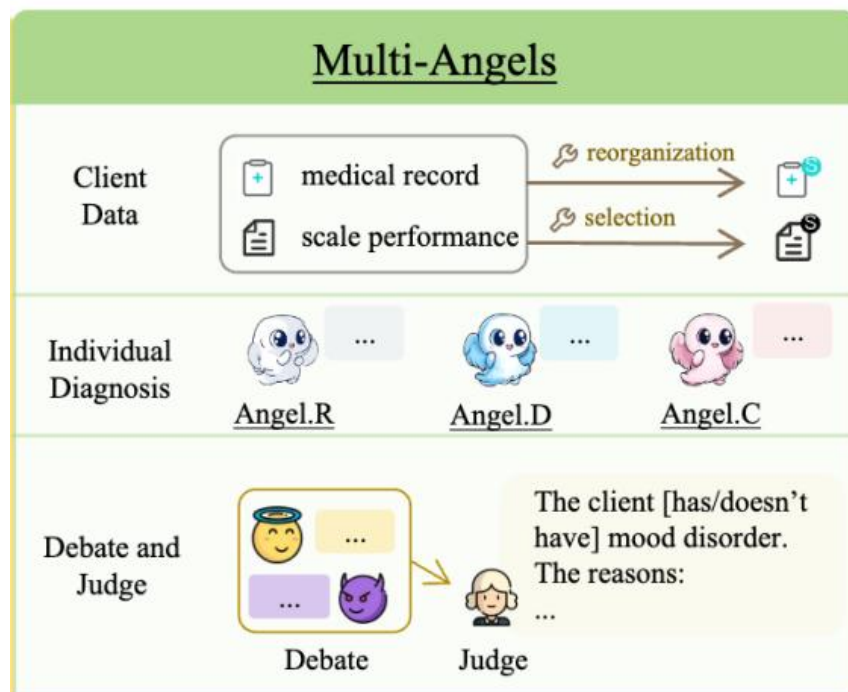
Framework Design: Three-Layer Core Mechanism



- Instead of relying on total scale scores, decompose 13 clinical scales (8 self-reported + 5 clinician-rated) into item-level data;
- Use Pearson correlation to select the Top 5% most mood disorder-relevant questions (16 core items) from the scales, clustered into 5 diagnostic groups: "Depressive Mood", "Suicidal Ideation", "Energy/Interest Loss", "Anxiety", "Insomnia";
- Resolve "conflicts between self-reports and clinical evaluations" through consistency checks within groups.



Framework Design: Three-Layer Core Mechanism



Single-Agent Variants

Three independent diagnostic roles balancing historical experience and individual variability.



Angel.R: No reference to past cases; based solely on DSM-5, (optional) medical record and scale data;

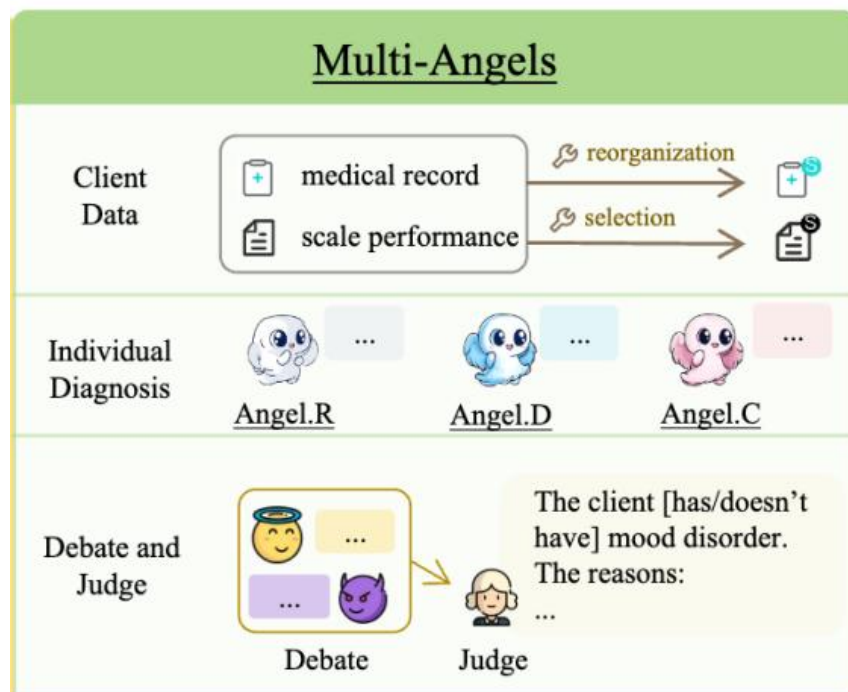


Angel.D: References Top-5 similar cases;



Angel.C: Compares and analyzes similarities / differences of similar cases.

Framework Design: Three-Layer Core Mechanism



Multi-Agent Debate

The Judge Agent integrates conclusions from the three single agents. In case of disagreements, a "Pro-Side vs. Con-Side" debate is triggered until a consistent diagnosis is reached.

A debate example:



Angel.R

The client has a mood disorder.

- The client shows mood instability with shifts from depression to (hypo)mania, consistent with bipolar disorder.
- Past hospitalization for mood disorder and long-term use of mood stabilizers support this.
- Symptoms like irritability, increased activity, and energy align with hypomanic episodes per DSM-5.



Angel.D

The client has a mood disorder.

- The client exhibits emotional instability, including hallucinations and inappropriate laughter.
- Despite clinician-rated scales showing no depression, self-reports suggest moderate depression.
- Similar cases were diagnosed as bipolar or major depression, reinforcing this possibility.



Angel.C

The client doesn't have a mood disorder.

- Though the client self-reports mild depressive and anxiety symptoms, clinician ratings show no signs of depression, anxiety, or insomnia.
- Mood fluctuation history is noted, but current assessments do not support an active mood disorder diagnosis.
- For hallucinations and inappropriate laughter, consider the client having schizophrenia.

A debate example:



Round 1

Both sides should claim the diagnosis and reasons in this round.



I believe this client has a mood disorder.

- The client's records show long-term mood fluctuations, ranging from depressive to elevated states, which is characteristic of bipolar disorder.
- The client was previously hospitalized for a mood disorder and has been on mood-related medication, indicating a strong link to mood disorder symptoms.
- Although clinician-rated assessments did not indicate significant depression or anxiety, the client's self-report showed moderate depression, which may better reflect their subjective emotional state.



I don't think this client has a mood disorder.

- Clinician-rated assessments show no signs of depression, anxiety, or suicidal tendencies—key indicators of mood disorders.
- Although the client initially improved with mood-related medication, new symptoms appeared despite continued medication use, suggesting a possible different mental condition.
- Similar cases have inconsistent diagnoses, so relying solely on past hospitalization records is insufficient for a conclusive diagnosis.



Rebuttals missing. Continue.

A debate example:



Round 2

Both sides should refute the opposite side.



- Self-reported symptoms indicate depression and suicidality, with a PHQ-9 score of 19 and daily suicidal ideation. These are core indicators of mood disorders, even if clinician-rated scales show minimal symptoms.
- New symptoms during continuous medication do not rule out mood disorders. Mood disorders often present episodically or with partial response to treatment.
- Historical diagnosis and treatment response support a mood disorder. The client was previously hospitalized for an affective disorder and showed improvement with mood stabilizers.



- Mood fluctuations alone do not confirm a mood disorder. The observed shifts could be part of other psychiatric conditions such as borderline personality traits or stress-induced responses. Symptom variability requires a broader differential diagnosis.
- Past hospitalization and medication use do not confirm the current diagnosis. Historical treatment for a mood-related condition is not definitive evidence of a present mood disorder, especially when recent clinician-rated assessments show no active depressive or anxious symptoms.
- Self-report scores must be interpreted cautiously. While PHQ-9 indicates moderate depression, clinician-administered scales (HAMD, HAMA, BPRS) consistently show no depressive, anxious, or suicidal signs. Clinical observation remains the gold standard.



Certain points are not discussed. Continue.

A debate example:



Round 3

Please discuss the possibility of a performance conflict between self-report and clinician assessment, in one sentence.



The client's symptoms are suppressed by medicine.



The client may have schizophrenia.



All symptoms discussed. Debate ends.

The client doesn't have a mood disorder.

Although the client was previously hospitalized and prescribed medication related to mood disorders, multiple clinician-administered scales did not indicate typical symptoms of depression or anxiety. Moreover, the client's current symptoms are not alleviated by mood disorder medications, raising the possibility of other psychiatric conditions with overlapping features, such as schizophrenia.



MoodSyn

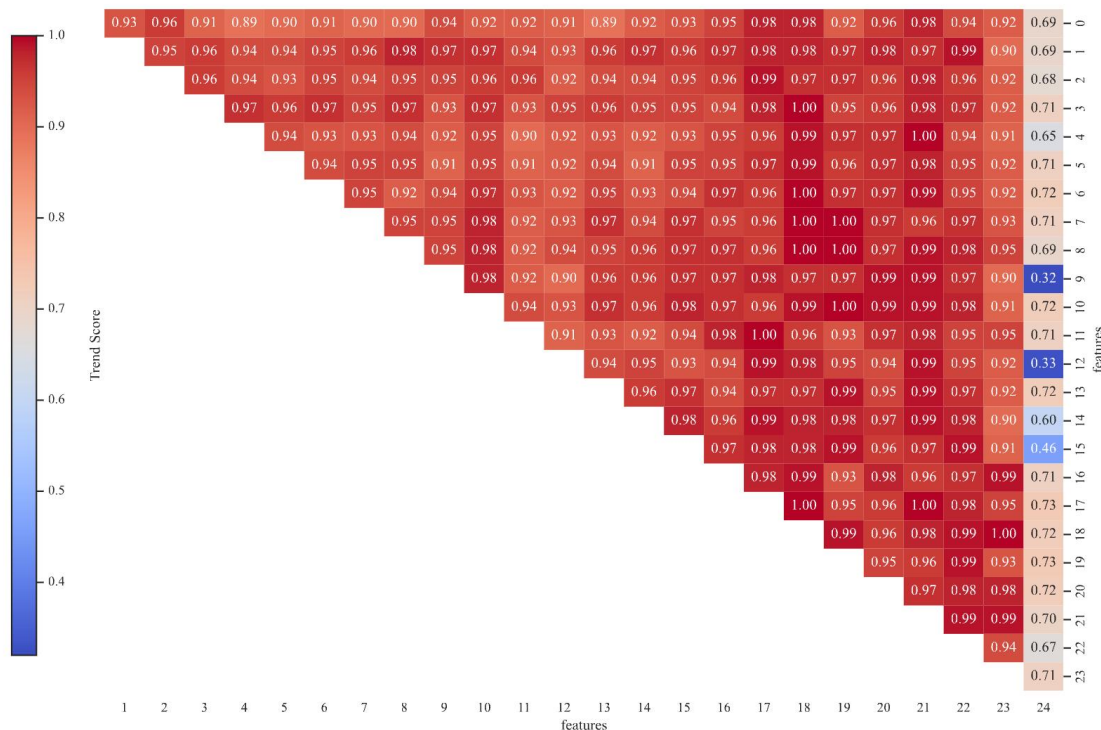
Privacy-Preserving Synthetic Clinical Dataset



```
{  
  "HAMA Q4 Score": 3,  
  "HAMA Q6 Score": 3,  
  "HAMA Total Score": 28,  
  "GAD7 Total Score": 18,  
  "PHQ9 Q1 Score": 3,  
  "PHQ9 Q2 Score": 2,  
  "PHQ9 Q4 Score": 2,  
  "PHQ9 Q9 Score": 0,  
  "PHQ9 Total Score": 14,  
  "HAMD Q1 Score": 2,  
  "HAMD Q3 Score": 1,  
  "HAMD Q4 Score": 1,  
  "HAMD Q7 Score": 2,  
  "HAMD Q22 Score": 1,  
  "HAMD Total Score": 28,  
  "BPRS Q9 Score": 3,  
  "PSQI Total Score": 11,  
  "SHAPS Total Score": 29,  
  "HCL32 Total Score": 18,  
  "DAS Total Score": 122,  
  "SSRS Total Score": 40,  
  "MDQ Total Score": 11,  
  "BPRS Total Score": 34,  
  "YMRS Total Score": 5,  
  "Mood Disorder": 1  
}
```

Dataset Specifications

- Scale: 1,173 synthetic cases, including a retrieval set (1,106 cases) and a test set (140 cases);
- Features: 25 core dimensions, including scores of 16 diagnostic items, total scores of 8 scales, and expert-annotated mood disorder labels;
- Generation Method: Based on the TabSyn framework (VAE + diffusion model), ensuring "statistical fidelity" and "clinical validity".

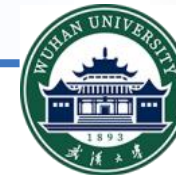


Core Values

- Privacy Security: No real patient information, avoiding data leakage risks;
- High Fidelity: Reproduces univariate / multivariate distributions of real data (density score = 0.86), and ML model training performance is close to that of real data (XGBoost classification ACC = 0.92);
- Open Access: Provides scarce high-quality training data for computational psychiatry research.

Key Experimental Results

Performance Surpassing Baselines



Model/Method	Real-Case ACC	Real-Case Recall	Comparative Advantage
GPT-4o (Baseline)	0.80	0.63	——
Angel.R (Basic Version)	0.92	0.84	12.3% higher ACC than GPT-4o
multi-Angels (Full Version)	0.925	0.881	Outperforms all single agents; significant improvement on hard cases
multi-Angels on MoodSyn	0.821	0.824	Maintains high robustness on synthetic data

Note: Metrics include Accuracy, Recall, MCC, and Macro F1, all outperforming 5 baseline models (e.g., LLaMA3, Mistral).

Conclusions & Research Significance

- **Technical Breakthrough:** MoodAngels is the first to realize psychiatric diagnosis with "retrieval augmentation + multi-agent debate", solving the problems of "subjective assessment bias" and "symptom overlap";
- **Data Contribution:** MoodSyn fills the gap in "privacy-preserving yet clinically valid" synthetic data, promoting reproducible research of AI in mental health;
- **Clinical Value:** Provides clinicians with an auxiliary diagnostic tool, especially suitable for scenarios with insufficient primary medical resources, facilitating early screening of mood disorders.



Angel.R



Angel.D



Angel.C



Judge



Positive



Negative



NEURAL INFORMATION
PROCESSING SYSTEMS



武汉大学
WUHAN UNIVERSITY

Thanks

Presenter: Mengxi Xiao

2025.10.24