

MAGI: Multi-Agent Guided Interview for Psychiatric Assessment

Guanqun Bi¹, Zhuang Chen^{2*}, Zhoufu Liu³, Hongkai Wang⁴,
Xiyao Xiao⁵, Yuqiang Xie⁶, Wen Zhang⁷, Yongkang Huang⁵,
Yuxuan Chen¹, Libiao Peng⁵, Minlie Huang¹

¹CoAI Group, DCST, IAI, BNRIST, Tsinghua University, ²Central South University,

³Beijing Institute of Technology, ⁴Sichuan University, ⁵Lingxin AI,

⁶Independent Researcher, ⁷University of International Relations

biguanqun@mail.tsinghua.edu.cn aihuang@tsinghua.edu.cn

Abstract

Automating structured psychiatric interviews could revolutionize mental healthcare accessibility, yet existing large language models (LLMs) approaches fail to align with psychiatric diagnostic protocols. We present MAGI, the first framework that transforms the gold-standard Mini International Neuropsychiatric Interview (MINI) into automatic computational workflows through coordinated multi-agent collaboration. MAGI dynamically navigates clinical logic via four specialized agents: 1) an interview tree guided navigation agent adhering to the MINI’s branching structure, 2) an adaptive question agent blending diagnostic probing, explaining, and empathy, 3) a judgment agent validating whether the response from participants meet the node, and 4) a diagnosis agent generating Psychometric Chain-of-Thought (PsyCoT) traces that explicitly map symptoms to clinical criteria. Experimental results on 1,002 real-world participants covering depression, generalized anxiety, social anxiety, and suicide show that MAGI advances LLM-assisted mental health assessment by combining clinical rigor, conversational adaptability, and explainable reasoning.

1 Introduction

Mental health disorders pose a significant global challenge, and accurate and timely diagnosis is crucial for effective intervention (American Psychiatric Association, 2013). Clinical interviews, particularly structured instruments, are gold standards for psychiatric assessment (Crisp and Chen, 2014). The Mini International Neuropsychiatric Interview (MINI) (Meyer et al., 2001) is a widely adopted structured interview protocol, providing a systematic approach to assess major psychiatric disorders. However, their manual administration is time-consuming and resource-intensive, creating a

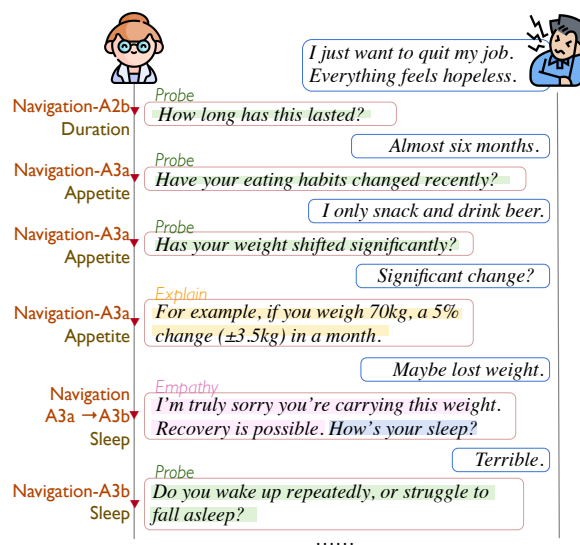


Figure 1: Example dialogue flow from MAGI. Our multi-agent framework MAGI guides participants through structured psychiatric interviews following MINI protocol.

bottleneck in mental healthcare access and limiting the widespread use in real-world settings.

The advent of large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023; DeepSeek-AI, 2024, 2025) offers promising avenues to automate and enhance mental health support processes, potentially democratizing access to care (He et al., 2023; Zhao et al., 2023). While LLM-based mental health tools demonstrate conversational fluency (Tu et al., 2024; Sartori and Orrù, 2023), three fundamental mismatches with clinical practice impede their adoption. **(1) Diagnostic framework deviation:** Current open-ended LLM dialogues frequently diverge from established clinical standards like DSM-5, particularly struggling with nuanced differential diagnoses (American Psychiatric Association, 2013). **(2) Rigid interaction patterns:** Existing systems employ inflexible questioning templates that prove inadequate when participants exhibit emotional distress or require clarifications,

*Corresponding author.

contradicting evidence-based interviewing principles (Milintsevich et al., 2023; Zhang and Guo, 2024). (3) **Limited diagnosis transparency:** Most systems operate as “black boxes”, providing little insight into their diagnostic reasoning process. This opacity significantly undermines trust and hinders clinical adoption, especially in high-stakes healthcare settings (Sadeghi et al., 2023).

To bridge these gaps, we present MAGI (Multi-Agent Guided Interview), a clinical logic-driven framework that operationalizes the MINI interview through four coordinated agents. Consider a depressive episode screening shown in Figure 1: when a participant reports feeling hopeless, the navigation agent activates the diagnostic nodes systematically from duration (A2b) to appetite (A3a) and sleep (A3b), dynamically managing interview progression. Guided by this structure, the question agent adaptively switches between three strategies: structure-driven probing (“How long has this lasted?”), explaining for ambiguous concepts (clarifying significant weight change), and empathetic support when detecting distress (Sharma et al., 2023). The judgment agent continuously validates responses against MINI criteria, iterating until all necessary evidence is collected. Finally, the diagnosis agent synthesizes the accumulated evidence into formal diagnostic conclusions through Psychometric Chain-of-Thought (PsyCoT), a clinical reasoning paradigm that explicitly connects symptom observations to diagnostic criteria via intermediate psychiatric constructs.

Through a school-based study, we collect 1,002 clinical interviews covering depression, generalized anxiety, social anxiety, and suicide risk. With dual-expert annotation from licensed psychologists, our experiments demonstrate MAGI’s clinical validity, achieving high agreement with expert diagnoses while providing transparent diagnostic traces absent in prior LLM approaches (Galatzer-Levy et al., 2023; Chen et al., 2023).

Our contributions are threefold:

(1) **Clinical-Grounded Architecture:** We present a novel multi-agent framework MAGI that effectively embodies the procedural logic of the MINI within a dynamic semi-structured dialogue system, advancing the state-of-the-art in automated clinical interviews.

(2) **Explainable Diagnostic Reasoning:** We propose the Psychometric Chain-of-Thought (PsyCoT) reasoning paradigm, which enhances the transparency and clinical interpretability of AI-driven

diagnostic processes by explicitly mirroring psychological measurement principles.

(3) **Large-Scale Clinical Validation:** Experiments through a school-based study with 1,002 participants and dual-expert clinical annotation demonstrates the framework’s superior diagnostic performance against baseline methods.

2 Related Work

2.1 Automatic Mental Health Assessment

Traditional psychiatric assessment methods, including standardized questionnaires and expert interviews (Meyer et al., 2001), face engagement and accessibility challenges (Merry et al., 2012). To address these limitations, researchers have explored computerized adaptive testing (Meijer and Nering, 1999) and game-based electronic assessments (Jones, 1984; Kim et al., 2016) to facilitate intrinsic motivation while maintaining assessment validity. Recent natural language processing advances have enabled sophisticated analysis of mental health indicators (Sharma et al., 2020, 2023), while LLMs have demonstrated potential in cognitive distortion detection (Chen et al., 2023), psychologist role-playing (Tu et al., 2024), and clinical interview analysis (Galatzer-Levy et al., 2023).

LLM-based multi-agent systems have proven effective in complex tasks requiring coordination (Park et al., 2023). However, existing LLM-based psychological assessment approaches often struggle with generalizability across psychological constructs (Chen et al., 2023) and maintaining consistent engagement (Tu et al., 2024). Our work introduces a novel multi-agent framework that combines gamification’s engagement benefits with LLMs’ flexibility and capabilities, addressing both generalizability across psychological constructs and participant engagement through interactive narrative generation.

2.2 MINI: A Structured Clinical Interview

The Mini International Neuropsychiatric Interview (MINI) is a structured diagnostic interview widely used for assessing major psychiatric disorders (Sheehan et al., 1998a). Aligned with DSM-5 criteria (American Psychiatric Association, 2013), it provides a brief yet comprehensive assessment format that ensures consistency across different interviewers (Lecrubier et al., 1997).

The MINI implements a decision tree logic where patient responses guide the interview pro-

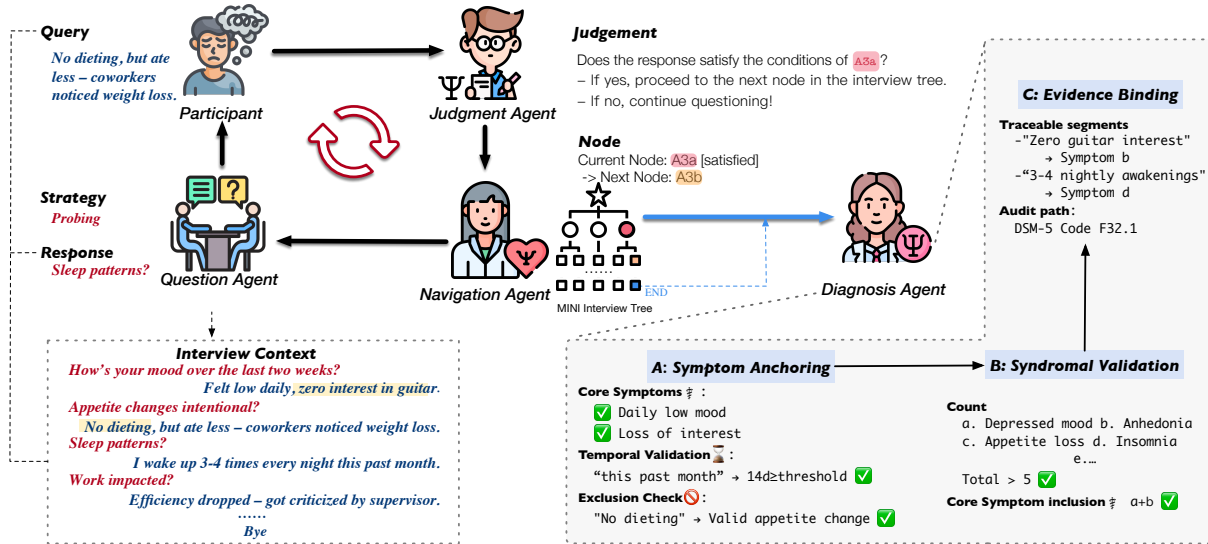


Figure 2: Overview of MAGI framework. The framework consists of a navigation agent for interview management, a question agent for dynamic utterance generation, a judgment agent for symptom validity analysis, and a diagnosis agent for DSM-5 compliant conclusions, ensuring adherence to psychiatric protocols and conversational adaptability.

gression (Amorim et al., 2019), ensuring standardized administration while maintaining comprehensive evaluation (Pettersson et al., 2018). Notable for its efficiency, the interview can be completed in 15-30 minutes without compromising diagnostic accuracy (Sheehan et al., 1998b). The instrument’s reliability and validity have been validated across diverse populations (Pinninti et al., 2003; Sheehan et al., 2010), establishing it as a psychiatric evaluation standard (van Vliet and de Beurs, 2015). Its standardized structure facilitates research comparability (Lecrubier et al., 1997) and presents opportunities for computational implementation (Malhotra and Chakrabarti, 2015). Our work adapts the MINI’s framework into a multi-agent collaborative system while preserving its diagnostic rigor in an automated format.

3 Methodology

MAGI transforms the MINI interview into a collaborative multi-agent process, where four specialized components interact through clinically constrained data flows. The navigation agent (§3.1) governs interview progression by enforcing MINI’s branching logic, while the question agent (§3.2) dynamically generates utterances that balance diagnostic probing with participant support. Participant responses are analyzed by the judgment agent (§3.3) to determine symptom validity, with accumulated evidence processed by the diagnosis agent (§3.4) to generate DSM-5 compliant conclusions through

structured reasoning traces. This architecture ensures adherence to psychiatric assessment protocols while maintaining conversational adaptability.

3.1 Navigation Agent

The navigation agent maintains overall control of the dialogue, enforcing strict adherence to the MINI’s semi-structured format. In practice, the agent governs dialog progression through hierarchical traversal of the MINI’s interview tree, where node transitions follow strict branching logic derived from diagnostic protocols. For instance, if a participant reports any sign of suicidal ideation, the agent routes the conversation to a deeper self-harm risk assessment node, whereas negative indications revert the session to a broader screening path. This design ensures that crucial items cannot be skipped or prematurely terminated: even if a participant attempts to change the topic, the agent compels the conversation to remain on the appropriate node until the relevant question set is satisfactorily answered.

By separating flow control from language generation, the system prevents “hallucinations” or creative detours that might undermine diagnostic integrity. This agent functions as a management center for MAGI. It coordinates with the judgment agent to force-choice constraints, ensuring that critical responses (e.g., indicators of suicidal thoughts or mania) are explicitly addressed. When incoming judgment indicate that a participant’s reply is ambiguous or shows reluctance to disclose infor-

mation, the navigation agent automatically directs the question agent to rephrase or clarify queries. This process upholds the core MINI principle that certain inquiries must not be bypassed or left unaddressed.

3.2 Question Agent

The question agent diversifies question phrasing to sustain engagement while adhering to diagnostic intent. For symptom assessment, it rephrases clinical constructs into conversational probes without deviating from diagnostic intent. For example, when probing for anhedonia — a core depression symptom — it adapts wording based on conversation history: a participant mentioning hobbies might hear "Have your favorite activities felt less rewarding recently?", while someone discussing social habits could receive "Do you still look forward to spending time with friends like before?". This contextual rephrasing prevents robotic repetition without compromising screening accuracy. When the LLM detects confusion through contextual inconsistencies (e.g., conflating chronic anxiety with momentary stress), it activates explanations: "Earlier you mentioned constant worry about work. This daily pattern aligns with what we call generalized anxiety.". Real-time dialogue analysis further identifies emotional distress cues through lexical patterns (e.g., repeated "overwhelmed", abrupt topic shifts), triggering empathetic support that maintains clinical progression: "I appreciate you sharing this. To help clarify, how many days last week did this loneliness feel most intense?". Strategy shifts emerge organically from LLM-powered context parsing rather than rigid hierarchies, enabling fluid transitions between diagnostic precision and human-centered responsiveness.

3.3 Judgment Agent

This agent operates real-time alignment between natural dialogue and structured diagnostic evaluation. It continuously evaluates response adequacy against predefined MINI criteria through three decision thresholds: direct matching of operationalized definitions (e.g., interpreting "I wake up panicking daily" as meeting generalized anxiety's physiological symptom threshold), semantic comprehension of clinically equivalent expressions ("constant worrying" → generalized anxiety criteria), and ambiguity resolution after 5 unproductive turns. Unmet criteria trigger recursive clarification requests through the question agent, while confirmed matches autho-

rize state transitions via the navigation agent. The forced-choice mechanism activates upon persistent ambiguity, presenting binary MINI-compliant options ("Would you describe this as [exact phrasing from MINI] or [alternative]?") to maintain diagnostic fidelity. This dual-layer validation - combining dynamic interpretation with protocol enforcement - ensures structured evidence collection while preserving natural dialogue flow.

3.4 Diagnosis Agent

The diagnosis agent operationalizes the PsyCoT (Psychometric Chain-of-Thought) framework to synthesize clinical evidence into DSM-5 compliant diagnoses through three interpretable reasoning phases.

At the **symptom anchoring** stage, each diagnostic criterion (e.g., " ≥ 5 symptoms over two weeks" for depression, "pervasive worry for six months" for generalized anxiety) is translated into multi-step validation logic. This includes symptom existence confirmation, temporal verification, and exclusionary condition checks (e.g., ruling out medical causes). The process transforms textual criteria into auditable decision paths. **Diagnostic synthesis** then performs syndromal validation by confirming that the symptom profile satisfies DSM-5 requirements. This includes verifying that the total number of distinct symptoms exceeds the diagnostic threshold, that core symptoms. This structured validation ensures that both the quantity and quality of symptoms align with the formal criteria before proceeding to diagnosis. Finally, **evidence binding** creates transparent audit trails by linking diagnostic conclusions to originating dialogue evidence and verification steps. Each judgment documents: which responses confirmed criteria fulfillment, how duration/exclusion rules were applied, and uncertainty sources requiring clinical review.

This three-phase PsyCoT implementation achieves transparency while maintaining MINI's diagnostic rigor, enabling clinicians to trace how conversational fragments map to specific DSM-5 codes through the reasoning chain.

4 Experiments

Data Collection We evaluated MAGI on 1,002 high-quality interview sessions collected through university mental health services. The interface for data collection is displayed in the Appendix B. Considering ethical constraints, we first construct sim-

Method	Rel.	Acc.	Comp.	Gui.
Direct	3.506	3.463	3.350	3.294
Role-Play	3.438	3.438	3.331	3.244
Knowledge	3.544	3.506	3.425	3.325
MINI	3.519	3.475	3.369	3.313
MAGI	3.576	3.503	3.497	3.460

Table 1: Results of dialogue quality evaluation. Rel.: Relevance, Acc.: Accuracy, Comp.: Completeness, Gui.: Guidance.

ulated participants with different mental illnesses based on LLMs. Then, we let various interviewers engage in conversations with them, and finally, we evaluate the interviews to compare the performance of different interviewers. Diagnostic annotations were performed by two licensed psychologists from APA-accredited university clinics, with 10-15 years of clinical experience. Following institutional compensation standards of \$100/hour, these experts implemented a three-phase annotation protocol. First, they jointly annotated 150 cases, achieving strong inter-rater reliability with an ICC of 0.78 and mean symptom dimension agreement of 0.69. They then independently annotated the remaining cases with weekly calibration sessions, maintaining high agreement (Cohen’s $\kappa > 0.85$). A final random re-sampling of 50 cases confirmed the annotation quality with an ICC of 0.87.

4.1 Dialogue Quality Evaluation

4.1.1 Experimental Setup

Baselines We evaluate four baseline approaches for psychiatric interview simulation: (1) *Direct*: The LLM follows explicit instructions to conduct diagnostic interviews without role-specific prompting or domain knowledge integration. (2) *Role-Play*: The LLM is instructed to simulate clinician expertise through the prompt "Act as a board-certified psychiatrist" while maintaining standard conversational capabilities. (3) *Knowledge*: The model is enhanced with DSM-5 diagnostic criteria and psychopathology knowledge graphs, enabling disorder-specific symptom evaluation during dialogues. (4) *MINI*: The LLM implements the MINI decision-tree architecture through constrained response generation and diagnostic branching logic.

Metrics For evaluation, experts rated each dialogue on 5-point Likert scales for: (1) Relevance:

Measures response alignment with participant’s queries and emotional context; (2) Accuracy: Evaluates psychological principle adherence and emotion judgment correctness; (3) Completeness: Assesses comprehensive coverage of critical issues with logical structure; (4) Guidance: Examines proactive conversation steering and emotional exploration facilitation. The detailed scale guidebook is given in Appendix C.

4.1.2 Results

The evaluation results in Table 1 demonstrate that MAGI achieves superior performance in almost all key dimensions. In relevance, our multi-agent architecture ensures systematic adherence to MINI nodes through the interview tree in the navigation agent, dynamically filtering off-topic responses while maintaining natural transitions between diagnostic topics. In completeness, the judgment agent’s forced conclusive symptom matching and the navigation agent’s state-aware questioning strategy work synergistically to exhaustively cover required diagnostic criteria, addressing the incompleteness of rigid MINI-based baselines. The guidance score benefits from the questioning agent’s hybrid strategy where empathetic feedback stabilizes participant engagement. Although MAGI slightly trails the knowledge baseline in accuracy, this marginal gap likely stems from our intentional trade-off between strict symptom classification and participant-centered adaptability. The judgment agent occasionally prioritizes participant comfort over forcing definitive answers, whereas the knowledge baseline’s static disease definitions enable more deterministic scoring. Crucially, MAGI’s diagnostic agent compensates for this through DSM-5-aligned reasoning chains, ensuring clinically valid interpretations despite minor accuracy variations in individual responses. These results collectively validate that MAGI’s orchestration of structured protocol adherence, adaptive interaction, and expert-knowledge integration achieves human-like competency in psychological interviews. Detailed results can be found in the Appendix A.1.

4.2 Diagnostic Performance Evaluation

4.2.1 Experimental Setup

Baselines We evaluate diagnostic results on 1,002 expert-annotated samples using four LLMs: GPT-4o, Claude-3.5-sonnet, GLM-Zero, and DeepSeek-R1, with temperature set to 0.7 for all models. Each model runs in three configurations:

Disorder	Metric	GPT-4o			Claude-3.5-sonnet			GLM-Zero			DeepSeek-R1		
		Vanilla	CoT	PsyCoT	Vanilla	CoT	PsyCoT	Vanilla	CoT	PsyCoT	Vanilla	CoT	PsyCoT
Depression	Control F1	0.930	0.938	0.976	0.925	0.925	0.978	0.877	0.887	0.968	0.912	0.914	0.977
	Case F1	0.396	0.429	0.558	0.393	0.396	0.632	0.333	0.356	0.576	0.408	0.432	0.639
	Macro-F1	0.663	0.683	0.767	0.659	0.661	0.805	0.605	0.622	0.772	0.660	0.673	0.808
	Accuracy	0.875	0.888	0.954	0.867	0.866	0.958	0.792	0.808	0.941	0.846	0.850	0.957
	κ	0.337	0.374	0.535	0.331	0.335	0.610	0.256	0.282	0.544	0.343	0.370	0.616
Generalized Anxiety	Control F1	0.974	0.974	0.973	0.970	0.972	0.980	0.931	0.939	0.985	0.969	0.974	0.990
	Case F1	0.866	0.868	0.857	0.845	0.860	0.899	0.756	0.778	0.929	0.851	0.872	0.952
	Macro-F1	0.920	0.921	0.915	0.907	0.916	0.940	0.843	0.858	0.957	0.910	0.923	0.971
	Accuracy	0.957	0.957	0.955	0.949	0.953	0.967	0.892	0.904	0.975	0.949	0.957	0.983
	κ	0.841	0.842	0.831	0.815	0.832	0.880	0.690	0.719	0.914	0.821	0.847	0.941
Social Anxiety	Control F1	0.984	0.982	0.992	0.974	0.976	0.991	0.948	0.951	0.995	0.974	0.974	0.996
	Case F1	0.752	0.736	0.800	0.667	0.686	0.786	0.505	0.519	0.898	0.667	0.667	0.918
	Macro-F1	0.868	0.859	0.896	0.820	0.831	0.888	0.727	0.735	0.946	0.820	0.820	0.957
	Accuracy	0.969	0.967	0.984	0.952	0.956	0.982	0.906	0.911	0.990	0.952	0.952	0.992
	κ	0.736	0.719	0.792	0.644	0.665	0.777	0.467	0.482	0.893	0.644	0.644	0.914
Suicide Risk	Control F1	0.965	0.963	0.995	0.963	0.966	0.988	0.941	0.937	0.995	0.965	0.966	0.995
	Case F1	0.385	0.340	0.942	0.430	0.454	0.852	0.316	0.326	0.944	0.396	0.425	0.947
	Macro-F1	0.675	0.651	0.968	0.697	0.710	0.920	0.629	0.631	0.969	0.681	0.695	0.971
	Accuracy	0.933	0.930	0.990	0.931	0.935	0.977	0.892	0.884	0.990	0.933	0.935	0.991
	κ	0.363	0.318	0.936	0.400	0.427	0.839	0.259	0.262	0.938	0.373	0.401	0.942

Table 2: Diagnostic performance evaluation results. PsyCoT is our proposed method. Vanilla represents end-to-end prediction from transcripts. CoT is free-form reasoning before diagnosis.

- (1) Vanilla: End-to-end prediction from transcripts;
- (2) CoT: Free-form reasoning before diagnosis. (c)
- PsyCoT: Our DSM-5-aligned structured reasoning.

Metrics Following previous studies (Burdizzo et al., 2023; Chen et al., 2024), we evaluate model performance using: (1) macro-F1 within case/control groups, (2) overall macro-F1 and accuracy across all data, and (3) Cohen’s κ for agreement with expert annotations.

4.2.2 Results

The evaluation in Table 2 demonstrates PsyCoT’s effectiveness in enhancing LLM-based psychiatric assessment across diagnostic complexity levels. The framework consistently outperforms baseline methods, particularly in high-stakes scenarios where clinical precision matters most. For suicide risk detection, a task requiring nuanced differentiation between transient distress and pathological patterns, PsyCoT elevates κ to 0.839-0.942, achieving clinical acceptability with κ exceeding 0.6, while baseline methods fall short with values between 0.259 and 0.427. This improvement stems from PsyCoT’s symptom anchoring mechanism, which maps symptom mentions to DSM-5 thresholds while filtering confounding factors. PsyCoT also mitigates model capability disparities: smaller

models like GLM-Zero approach GPT-4o’s performance in depression screening with Macro-F1 scores of 0.772 versus 0.767 when guided by structured clinical reasoning. The framework balances sensitivity-specificity tradeoffs by maintaining control group F1 above 0.968 while doubling case detection rates for social anxiety disorders. This reflects its syndromic validation phase, which evaluates symptom interdependencies to reject isolated signals, such as distinguishing pathological avoidance from situational nervousness. The multi-agent architecture preserves MINI’s procedural rigor, as evidenced by DeepSeek-R1’s improved adherence to diagnostic hierarchies under PsyCoT, reducing misclassification of subclinical cases. These results suggest that encoding domain-specific reasoning chains compensates for LLMs’ tendency toward heuristic judgments, bridging the gap between open-ended language understanding and standardized diagnostic protocols.

5 Discussion and Analysis

5.1 The Effect of Few-shot Learning

Results in Table 3 demonstrate the superior performance of PsyCoT across four psychological disorders. In the zero-shot setting, PsyCoT achieves substantial improvements over baseline methods, par-

ticularly in suicide risk assessment with F1 scores of 0.945 and generalized anxiety detection with F1 scores of 0.946, indicating robust integration of clinical knowledge. The framework maintains consistent performance across disorders, with F1 scores ranging from 0.803 to 0.946 in zero-shot settings, suggesting effective incorporation of DSM-5 criteria and MINI protocols. While all methods benefit from few-shot learning, PsyCoT exhibits more stable improvements ranging from 1.5% to 5.9%, compared to the larger variations seen in baseline approaches—Vanilla methods show improvements from 2.7% to 26.3%, while CoT methods show improvements from 1.7% to 20.9%. This stability is particularly evident in challenging tasks such as depression assessment, where PsyCoT achieves a significant improvement from baseline with F1 increasing from 0.679 to 0.850. The most notable performance gain is observed in social anxiety detection, where the baseline method shows substantial few-shot improvement of 26.3%, indicating the complexity of social anxiety manifestations. These results validate our multi-agent architecture’s effectiveness in conducting structured psychological interviews, with PsyCoT demonstrating robust performance across different clinical scenarios, particularly in critical tasks such as suicide risk assessment with few-shot F1 reaching 0.963. The framework’s consistent performance across disorders and minimal reliance on few-shot examples suggests successful integration of clinical expertise into the reasoning process, supporting its potential for broad clinical application. Detailed few-shot methods results can be found in the Appendix A.2.

5.2 Interview Tree Nodes Analysis

Figure 3 presents heatmaps visualizing participant interaction patterns with different interview nodes. The color intensity represents normalized metrics ranging from 0 to 1. In the depression assessment, participants showed high interaction intensity at entry nodes a1a and a2a, while multi-turn interactions dominated middle-phase nodes such as ra3a, and force choice questions were prominent in later stages including ra3g and ra4. This progression suggests a structured approach from open dialogue to specific symptom confirmation. The generalized anxiety assessment demonstrated concentrated interaction at early nodes N1a and N2 with both entry and multi-turn patterns, indicating the importance of establishing rapport before detailed symptom exploration. Social anxiety evaluation exhibited

Disorder & Method		Macro-F1		
		Zero-Shot	Few-Shot	Δ (%)
Depression	Vanilla	0.679	0.723	+6.5
	CoT	0.700	0.755	+7.9
	PsyCoT	0.803	0.850	+5.9
Generalized Anxiety	Vanilla	0.895	0.919	+2.7
	CoT	0.904	0.919	+1.7
	PsyCoT	0.946	0.960	+1.5
Social Anxiety	Vanilla	0.729	0.921	+26.3
	CoT	0.761	0.920	+20.9
	PsyCoT	0.927	0.942	+1.6
Suicide	Vanilla	0.763	0.815	+6.8
	CoT	0.775	0.810	+4.5
	PsyCoT	0.945	0.963	+1.9

Table 3: Few-shot performance improvements across disorders with comparison of methods (Vanilla, CoT, PsyCoT). **Bold** values indicate best results from our proposed PsyCoT framework.

a unique pattern where entry interactions peaked at F1a and F1, followed by increased force choice and multi-turn interactions at nodes F4 through F5, highlighting the necessity of trust-building before exploring specific social scenarios. Suicide tendency assessment displayed more dispersed interaction patterns, with heightened intensity at nodes b17a and b18c, reflecting the complexity and sensitivity of this evaluation domain. The varying intensity patterns across different nodes and interaction types suggest that effective mental health assessment requires a dynamic approach, adapting the interaction style based on both the specific dimension being evaluated and the stage of assessment.

5.3 Case Study: Explainable Diagnosis

As shown in Table 4, the generalized anxiety case demonstrates MAGI’s capacity to transform ambiguous dialog exchanges into clinically valid diagnoses through transparent PsyCoT reasoning.

When the participant initially described intermittent worry patterns (*occur every few days*), the agent clarified temporal parameters through multi-turn verification, mathematically aggregating frequency reports to confirm DSM-5’s 6-month duration threshold. Core symptom validation involved distinguishing pathological uncontrollability (*find it very difficult to control my anxiety* despite self-help attempts) from transient stress through semantic decomposition of coping efficacy descriptions. Physiological markers were systematically mapped to diagnostic criteria: self-reported *stom-*

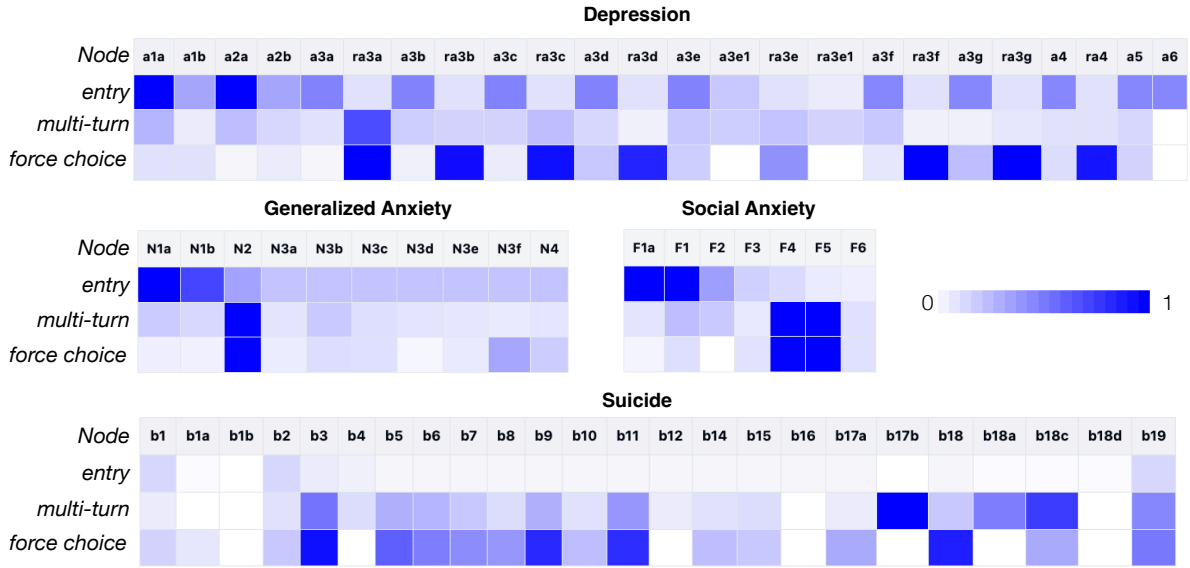


Figure 3: Heatmap illustrating participant interaction patterns with interview nodes across different disorders, with darker blue indicating higher percentage at different assessment nodes.

Case ID: GAD-xxxxxxx
Diagnosis: Generalized Anxiety (F41.1)
[Criterion A] Excessive worry confirmed via multi-turn temporal validation: <i>...occur every few days</i> → meets 6-month persistence through cumulative frequency analysis
[Criterion B] Uncontrollability established through coping efficacy assessment: <i>I find it very difficult to control my anxiety</i> → despite self-help attempts
[Criterion C] Physiological symptoms mapped through clinical equivalence: <i>Stomach spasms</i> → muscle tension <i>Sleep disturbance weighted as secondary evidence:</i> → frequent nightmares with fatigue
[Criterion E] Functional impairment validated: <i>Inappropriate attitudes during communication</i> → social domain impact

Table 4: Case study for PsyCoT. *Italic* words are the original dialogue from the interview. **Bolden** words are the analyzed symptoms.

ach spasms validated muscle tension through visceromotor equivalence reasoning, while sleep quality complaints (*frequent nightmares and daytime fatigue*) were weighted as secondary evidence under polythetic rules. Crucially, the agent maintained an audit trail linking each diagnostic conclusion to 11 dialog segments - for instance, binding social impairment to specific interpersonal conflict descriptions (*inappropriate attitudes during communication*) while excluding alternative explanations

through negative medical history screening. This end-to-end traceability, where conversational fragments flow through symptom anchoring (temporal verification), diagnostic synthesis (symptom constellation analysis), and evidence binding (DSM-5 criteria mapping), exemplifies how PsyCoT bridges natural language understanding with clinical decision protocols. Due to space limitations, we only show the most core part of the reasoning.

6 Conclusion

This work addresses the critical challenge of aligning LLM-powered mental health assessment with clinical diagnostic standards through MAGI, the first multi-agent framework that operationalizes structured psychiatric interviews. By translating the MINI protocol into dynamic computational workflows, our framework overcomes three fundamental limitations of existing approaches: (1) systematic adherence to diagnostic logic through navigation and judgment agents ensures clinical validity, (2) adaptive questioning strategies enable human-like conversational flexibility while maintaining assessment integrity, and (3) the novel PsyCoT mechanism provides transparent diagnostic reasoning that maps symptoms to DSM-5 criteria through intermediate psychiatric constructs. Extensive evaluation across 1,002 real-world cases demonstrates MAGI’s superior diagnostic agreement with expert clinicians while maintaining natural interaction flow.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (No. 2024YFC3606800), the National Science Foundation for Distinguished Young Scholars (No. 62125604), and the NSFC Project (No. 62441614). The authors thank all project team members for their dedicated work and valuable suggestions, and express gratitude to collaborating institutions for their strong support.

Ethical Considerations

Ethical Review and Informed Consent This study rigorously adheres to ethical standards through comprehensive measures spanning ethical oversight, psychological safeguards, and data governance. The experimental protocol received formal approval from an institutional review board. Prior to participation, all 1,002 individuals provided informed consent, explicitly acknowledging the non-diagnostic nature of the research, their unconditional right to withdraw without consequences, and the exclusive use of anonymized data for institutional management and academic publication. No biometric identifiers were collected or stored.

Psychological Risk Mitigation To address psychological risks, a multi-tiered intervention protocol was established for participants exhibiting signs of depression, anxiety, or suicidal ideation. Immediate referrals to certified psychological hotlines were provided during initial assessments, followed by secondary evaluations conducted by licensed clinicians within 72 hours. High-risk cases involving self-harm or suicidal intent triggered collaborative alerts between researchers and institutional authorities, with all procedures reviewed and validated by a panel of clinical psychology experts to align with the Ethical Code for Clinical and Counseling Psychology issued by the Chinese Psychological Society.

Data Management and AI Boundaries Data security was ensured through storage on a government-certified cloud platform meeting cybersecurity standards, accessible only to authorized research members. Findings were aggregated into non-identifiable group reports to preclude individual disclosure. Notably, the study enforced strict boundaries on AI applications: no AI-generated analyses were directly shared with participants,

clinical diagnoses were reserved exclusively for qualified psychiatrists, and no data were transferred to third-party commercial entities. These safeguards collectively prioritize participant welfare, data integrity, and compliance with regional ethical-legal frameworks.

Limitations and Future Work

The current implementation of MAGI faces several important constraints. While our evaluation demonstrates strong agreement with expert judgments, longitudinal studies with diverse participant populations are needed to validate MAGI's reliability in clinical settings. Technical aspects requiring further development include enhancement of the PsyCoT reasoning framework for complex comorbidity patterns and improvement of real-time emotion detection capabilities. Practical deployment considerations encompass clinical integration protocols, healthcare provider training requirements, and robust data privacy frameworks. These challenges, while significant, also present opportunities for advancing AI-assisted mental healthcare through careful attention to clinical, ethical, and technical considerations.

References

- American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, 5th edition. American Psychiatric Association, Washington, D.C.
- Patricia Amorim, Yves Lecrubier, Emmanuelle Weiller, Thierry Hergueta, and David Sheehan. 2019. Diagnostic accuracy of the mini-international neuropsychiatric interview in psychosis and bipolar disorder: A meta-analysis. *Comprehensive Psychiatry*, 89:44–52.
- Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. In *Proc. Interspeech 2023*, pages 3617–3621.
- Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.
- Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. 2024. [Depression detection in clinical interviews with llm-empowered structural element graph](#). In *North American Chapter of the Association for Computational Linguistics*.

- Nigel Crisp and Lincoln Chen. 2014. Global supply of health professionals. *New England Journal of Medicine*, 370(10):950–957.
- DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longevity value. *arXiv preprint arXiv:2401.02954*.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). Preprint, arXiv:2501.12948.
- Isaac R Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01834*.
- Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, et al. 2023. Towards a psychological generalist ai: A survey of current applications of large language models and future prospects. *arXiv preprint arXiv:2312.04578*.
- Marshall B Jones. 1984. Video games as psychological tests. *Simulation & Games*, 15(2):131–157.
- Yoon Jeon Kim, Russell G Almond, and Valerie J Shute. 2016. Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing*, 16(2):142–163.
- Y Lecrubier, DV Sheehan, E Weiller, P Amorim, I Bonora, KH Sheehan, J Janavs, and GC Dunbar. 1997. The mini international neuropsychiatric interview (mini). a short diagnostic structured interview: reliability and validity according to the cidi. *European Psychiatry*, 12(5):224–231.
- Sanjeev Malhotra and Subho Chakrabarti. 2015. Computerized diagnostic interviews for psychiatric disorders: A systematic review. *Indian Journal of Psychiatry*, 57(2):137.
- Rob R Meijer and Michael L Nering. 1999. Computerized adaptive testing: Overview and introduction. *Applied psychological measurement*, 23(3):187–194.
- Sally N Merry, Karolina Stasiak, Matthew Shepherd, Chris Frampton, Theresa Fleming, and Mathijs FG Lucassen. 2012. The effectiveness of sparx, a computerised self help intervention for adolescents seeking help for depression: randomised controlled non-inferiority trial. *Bmj*, 344.
- Gregory J Meyer, Stephen E Finn, Lorraine D Eyde, Gary G Kay, Kevin L Moreland, Robert R Dies, Elena J Eisman, Tom W Kubiszyn, and Geoffrey M Reed. 2001. Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, 56(2):128.
- Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1).
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Agneta Pettersson, Kristina B Boström, Petter Gustavsson, and Lisa Ekselius. 2018. The mini-international neuropsychiatric interview is useful and well accepted as part of the clinical assessment for depression and anxiety in primary care. *Family Practice*, 35(6):733–739.
- Narsimha R Pinninti, Heather Madison, Erica Musser, and David Rissmiller. 2003. The mini international neuropsychiatric interview: Reliability and validity in relation to scid. *European Psychiatry*, 18(7):374–377.
- Misha Sadeghi, Bernhard Egger, Reza Agahi, Robert Richer, Klara Capito, Lydia Helene Rupp, Lena Schindler-Gmelch, Matthias Berking, and Bjoern M Eskofier. 2023. Exploring the capabilities of a language model-only approach for depression detection in text data. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5. IEEE.
- Giuseppe Sartori and Graziella Orrù. 2023. Language models and psychological sciences. *Frontiers in Psychology*, 14.
- Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.
- Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000.
- David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, Geoffrey C Dunbar, et al. 1998a. The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *J clin psychiatry*, 59(Suppl 20):22–33.
- David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Juris Janavs, Emmanuelle Weiller, Arif Keskiner,

John Schinka, Emily Knapp, Michael F Sheehan, and Geoffrey C Dunbar. 1998b. Comparative validity of the mini international neuropsychiatric interview (mini) and the composite international diagnostic interview (cidi). *European Psychiatry*, 13(4):232–241.

David V Sheehan, Kathy H Sheehan, R Douglas Shytle, Juris Janavs, Yvonne Bannon, Jamison E Rogers, Karen M Milo, Sandra L Stock, and Berney Wilkinson. 2010. Reliability and validity of the mini international neuropsychiatric interview for children and adolescents (mini-kid). *The Journal of clinical psychiatry*, 71(3):17393.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Irene M van Vliet and Edwin de Beurs. 2015. Validation of the mini international neuropsychiatric interview (mini) as a screening tool for substance use disorders in a psychiatric hospital. *Addictive Behaviors*, 41:128–131.

Jun Zhang and Yanrong Guo. 2024. Multilevel depression status detection based on fine-grained prompt learning. *Pattern Recognition Letters*, 178:167–173.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Detailed Diagnosis Results

A.1 Zero-shot Results

The zero-shot evaluation reveals significant performance variations across psychological disorders and prompting strategies, as shown in Table 5. For depression diagnosis, PsyCoT consistently outperformed vanilla and chain-of-thought prompting, achieving Macro-F1 scores between 0.767 and 0.808 and Cohen’s κ values between 0.535 and 0.616 across models. GPT-4o with PsyCoT improved case F1 from 0.396 to 0.558, representing a 41% increase compared to vanilla prompting and demonstrating enhanced sensitivity to minority-class identification. This pattern extends to high-stakes domains: for suicide risk assessment, PsyCoT increased GPT-4o’s case F1 from 0.385 to

0.942, a 145% improvement that substantially reduced false negatives. Model-specific disparities emerged in the results, with Claude-3.5 showing stronger anxiety detection with a Macro-F1 of 0.940 using PsyCoT, while GLM-Zero benefited most in social anxiety classification with Macro-F1 improving from 0.946 to 0.975, a 29% increase. The κ values between 0.936 and 0.942 for suicide risk suggest PsyCoT’s improved clinical reliability by aligning with multi-rater consensus standards. A critical limitation persists: while control-group precision remains near-perfect with values between 0.962 and 0.997 across disorders, case recall fluctuates considerably, ranging from 0.439 to 0.966 for depression, indicating potential dataset bias toward majority classes. These results highlight task-specific optimization requirements and the necessity of domain-adapted prompting frameworks in clinical NLP applications.

A.2 Few-shot Results

The few-shot analysis demonstrates enhanced diagnostic capabilities through in-context learning, with PsyCoT achieving state-of-the-art performance across most metrics as shown in Table 6. For suicide risk assessment, a critical clinical task, PsyCoT improved GPT-4o’s case F1 from 0.500 to 0.927, an 85% increase, and Cohen’s κ from 0.475 to 0.920, a 93% improvement compared to vanilla prompting. This approach achieved near-perfect control group recall of 0.990 while maintaining high precision of 0.902. Model specialization emerged in the results: Claude-3.5 attained peak anxiety detection performance with a Macro-F1 of 0.934 using 3-shot PsyCoT, while GLM-Zero showed dramatic improvements in depression classification with Macro-F1 increasing by 19% to 0.788 and κ increasing by 72% to 0.580. DeepSeek-R1 achieved the highest social anxiety detection scores with a Macro-F1 of 0.973 through vanilla few-shot learning, suggesting certain architectures benefit more from direct exemplar exposure. However, PsyCoT universally enhanced inter-rater reliability, particularly for suicide risk where κ values exceeded 0.900 across all models—a threshold indicating clinical-grade consensus. The results reveal an inverse precision-recall tradeoff: while control group precision remained stable at 0.987 with a standard deviation of 0.011 across disorders, case recall showed greater variance, ranging from 0.485 to 0.954 for depression in PsyCoT. This highlights the need for task-specific calibration when

Disorder	Model	Method	Control			Case			Macro-F1	Acc.	κ
			P	R	F1	P	R	F1			
Depression	GPT-4o	Vanilla	0.971	0.893	0.930	0.291	0.621	0.396	0.663	0.875	0.337
		CoT	0.973	0.906	0.938	0.323	0.636	0.429	0.683	0.888	0.374
		PsyCoT	0.962	0.990	0.976	0.763	0.439	0.558	0.767	0.954	0.535
	Claude-3.5	Vanilla	0.973	0.882	0.925	0.281	0.652	0.393	0.659	0.867	0.331
		CoT	0.974	0.880	0.925	0.282	0.667	0.396	0.661	0.866	0.335
		PsyCoT	0.969	0.987	0.978	0.750	0.545	0.632	0.805	0.958	0.610
	GLM-Zero	Vanilla	0.981	0.793	0.877	0.211	0.788	0.333	0.605	0.792	0.256
		CoT	0.983	0.809	0.887	0.228	0.803	0.356	0.622	0.808	0.282
		PsyCoT	0.972	0.965	0.968	0.548	0.606	0.576	0.772	0.941	0.544
	DeepSeek-R1	Vanilla	0.984	0.849	0.912	0.273	0.803	0.408	0.660	0.846	0.343
		CoT	0.989	0.849	0.914	0.288	0.864	0.432	0.673	0.850	0.370
		PsyCoT	0.970	0.984	0.977	0.717	0.576	0.639	0.808	0.957	0.616
Generalized Anxiety	GPT-4o	Vanilla	0.953	0.996	0.974	0.979	0.777	0.866	0.920	0.957	0.841
		CoT	0.956	0.994	0.974	0.966	0.788	0.868	0.921	0.957	0.842
		PsyCoT	0.949	0.999	0.973	0.993	0.754	0.857	0.915	0.955	0.831
	Claude-3.5	Vanilla	0.953	0.987	0.970	0.927	0.777	0.845	0.907	0.949	0.815
		CoT	0.959	0.985	0.972	0.923	0.804	0.860	0.916	0.953	0.832
		PsyCoT	0.963	0.999	0.980	0.993	0.821	0.899	0.940	0.967	0.880
	GLM-Zero	Vanilla	0.984	0.883	0.931	0.635	0.933	0.756	0.843	0.892	0.690
		CoT	0.985	0.897	0.939	0.664	0.939	0.778	0.858	0.904	0.719
		PsyCoT	0.981	0.989	0.985	0.948	0.911	0.929	0.957	0.975	0.914
	DeepSeek-R1	Vanilla	0.961	0.978	0.969	0.890	0.816	0.851	0.910	0.949	0.821
		CoT	0.962	0.987	0.974	0.930	0.821	0.872	0.923	0.957	0.847
		PsyCoT	0.986	0.994	0.990	0.971	0.933	0.952	0.971	0.983	0.941
Social Anxiety	GPT-4o	Vanilla	0.999	0.969	0.984	0.610	0.979	0.752	0.868	0.969	0.736
		CoT	0.998	0.968	0.982	0.597	0.958	0.736	0.859	0.967	0.719
		PsyCoT	0.984	1.000	0.992	1.000	0.667	0.800	0.896	0.984	0.792
	Claude-3.5	Vanilla	1.000	0.950	0.974	0.500	1.000	0.667	0.820	0.952	0.644
		CoT	1.000	0.954	0.976	0.522	1.000	0.686	0.831	0.956	0.665
		PsyCoT	0.984	0.997	0.991	0.917	0.688	0.786	0.888	0.982	0.777
	GLM-Zero	Vanilla	1.000	0.901	0.948	0.338	1.000	0.505	0.727	0.906	0.467
		CoT	1.000	0.907	0.951	0.350	1.000	0.519	0.735	0.911	0.482
		PsyCoT	0.996	0.994	0.995	0.880	0.917	0.898	0.946	0.990	0.893
	DeepSeek-R1	Vanilla	1.000	0.950	0.974	0.500	1.000	0.667	0.820	0.952	0.644
		CoT	1.000	0.950	0.974	0.500	1.000	0.667	0.820	0.952	0.644
		PsyCoT	0.997	0.995	0.996	0.900	0.938	0.918	0.957	0.992	0.914
Suicide Risk	GPT-4o	Vanilla	0.933	0.999	0.965	0.955	0.241	0.385	0.675	0.933	0.363
		CoT	0.930	0.999	0.963	0.947	0.207	0.340	0.651	0.930	0.318
		PsyCoT	0.993	0.996	0.995	0.953	0.931	0.942	0.968	0.990	0.936
	Claude-3.5	Vanilla	0.937	0.991	0.963	0.765	0.299	0.430	0.697	0.931	0.400
		CoT	0.938	0.995	0.966	0.844	0.310	0.454	0.710	0.935	0.427
		PsyCoT	0.978	0.998	0.988	0.971	0.759	0.852	0.920	0.977	0.839
	GLM-Zero	Vanilla	0.933	0.950	0.941	0.352	0.287	0.316	0.629	0.892	0.259
		CoT	0.936	0.938	0.937	0.329	0.322	0.326	0.631	0.884	0.262
		PsyCoT	0.997	0.992	0.995	0.923	0.966	0.944	0.969	0.990	0.938
	DeepSeek-R1	Vanilla	0.934	0.998	0.965	0.917	0.253	0.396	0.681	0.933	0.373
		CoT	0.935	0.998	0.966	0.923	0.276	0.425	0.695	0.935	0.401
		PsyCoT	0.992	0.998	0.995	0.976	0.920	0.947	0.971	0.991	0.942

Table 5: Comparative performance analysis of LLMs with zero-shot prompting strategies.

deploying few-shot models in clinical practice.

B Demo

We introduce the demo of MAGI, which has been successfully deployed and widely adopted by a large user base. The chatbot offers questions in both text and audio formats, ensuring an inclusive experience for users with diverse needs. Through this interactive and intuitive interface, our system supports self-reflection and early mental health screening, positioning it as a valuable tool for large-scale implementation.

Real-world usage has shown that our system not only helps users articulate their emotions effectively but also aligns seamlessly with established diagnostic frameworks, demonstrating its potential to make a meaningful impact on mental health awareness and care.

Disorder	Model	Method	Control			Case			Macro-F1	Acc.	κ
			P	R	F1	P	R	F1			
Depression	GPT-4o	Vanilla	0.961	0.951	0.956	0.395	0.455	0.423	0.689	0.918	0.379
		CoT	0.962	0.957	0.960	0.437	0.470	0.453	0.706	0.925	0.412
		PsyCoT	0.965	0.991	0.978	0.800	0.485	0.604	0.791	0.958	0.583
	Claude-3.5	Vanilla	0.971	0.933	0.951	0.388	0.606	0.473	0.712	0.911	0.427
		CoT	0.970	0.948	0.959	0.443	0.591	0.506	0.733	0.924	0.466
		PsyCoT	0.967	0.994	0.980	0.850	0.515	0.642	0.811	0.962	0.623
	GLM-Zero	Vanilla	0.974	0.881	0.925	0.284	0.667	0.398	0.662	0.867	0.337
		CoT	0.975	0.878	0.924	0.283	0.682	0.400	0.662	0.865	0.338
		PsyCoT	0.988	0.938	0.962	0.487	0.833	0.615	0.788	0.931	0.580
	DeepSeek-R1	Vanilla	0.966	0.978	0.972	0.618	0.515	0.562	0.767	0.947	0.534
		CoT	0.969	0.973	0.971	0.597	0.561	0.578	0.775	0.946	0.549
		PsyCoT	0.969	0.989	0.979	0.783	0.545	0.643	0.811	0.960	0.622
Generalized Anxiety	GPT-4o	Vanilla	0.955	0.998	0.976	0.986	0.782	0.872	0.924	0.959	0.848
		CoT	0.956	0.996	0.976	0.979	0.788	0.873	0.924	0.959	0.849
		PsyCoT	0.975	1.000	0.987	1.000	0.883	0.938	0.963	0.979	0.925
	Claude-3.5	Vanilla	0.973	0.978	0.976	0.897	0.877	0.887	0.931	0.960	0.863
		CoT	0.971	0.977	0.974	0.891	0.866	0.878	0.926	0.957	0.852
		PsyCoT	0.959	0.999	0.979	0.993	0.804	0.889	0.934	0.964	0.868
	GLM-Zero	Vanilla	0.986	0.944	0.965	0.785	0.939	0.855	0.910	0.943	0.820
		CoT	0.987	0.948	0.967	0.797	0.944	0.864	0.916	0.947	0.832
		PsyCoT	0.991	0.984	0.988	0.930	0.961	0.945	0.966	0.980	0.933
	DeepSeek-R1	Vanilla	0.966	0.973	0.970	0.873	0.844	0.858	0.914	0.950	0.828
		CoT	0.961	0.978	0.969	0.890	0.816	0.851	0.910	0.949	0.821
		PsyCoT	0.989	0.994	0.992	0.971	0.950	0.960	0.976	0.986	0.952
Social Anxiety	GPT-4o	Vanilla	0.993	1.000	0.996	1.000	0.854	0.921	0.959	0.993	0.918
		CoT	0.992	1.000	0.996	1.000	0.833	0.909	0.952	0.992	0.905
		PsyCoT	0.988	1.000	0.994	1.000	0.750	0.857	0.925	0.988	0.851
	Claude-3.5	Vanilla	0.991	0.993	0.992	0.848	0.813	0.830	0.911	0.984	0.821
		CoT	0.992	0.992	0.992	0.833	0.833	0.833	0.912	0.984	0.825
		PsyCoT	0.987	0.999	0.993	0.972	0.729	0.833	0.913	0.986	0.826
	GLM-Zero	Vanilla	1.000	0.958	0.979	0.545	1.000	0.706	0.842	0.960	0.686
		CoT	1.000	0.965	0.982	0.593	1.000	0.744	0.863	0.967	0.728
		PsyCoT	0.995	0.996	0.995	0.915	0.896	0.905	0.950	0.991	0.901
	DeepSeek-R1	Vanilla	0.999	0.996	0.997	0.922	0.979	0.949	0.973	0.995	0.947
		CoT	0.997	0.994	0.995	0.882	0.938	0.909	0.952	0.991	0.904
		PsyCoT	0.998	0.999	0.998	0.979	0.958	0.968	0.983	0.997	0.967
Suicide Risk	GPT-4o	Vanilla	0.941	0.997	0.968	0.909	0.345	0.500	0.734	0.940	0.475
		CoT	0.940	0.997	0.968	0.906	0.333	0.487	0.728	0.939	0.462
		PsyCoT	0.996	0.990	0.993	0.902	0.954	0.927	0.960	0.987	0.920
	Claude-3.5	Vanilla	0.960	0.960	0.960	0.575	0.575	0.575	0.767	0.926	0.534
		CoT	0.948	0.958	0.953	0.506	0.448	0.476	0.714	0.914	0.429
		PsyCoT	0.990	0.993	0.992	0.929	0.897	0.912	0.952	0.985	0.904
	GLM-Zero	Vanilla	0.938	0.983	0.960	0.636	0.322	0.427	0.694	0.925	0.392
		CoT	0.938	0.984	0.960	0.643	0.310	0.419	0.689	0.925	0.384
		PsyCoT	0.997	0.993	0.995	0.933	0.966	0.949	0.972	0.991	0.944
	DeepSeek-R1	Vanilla	0.950	0.995	0.972	0.886	0.448	0.595	0.784	0.947	0.570
		CoT	0.944	0.990	0.966	0.786	0.379	0.512	0.739	0.937	0.482
		PsyCoT	0.991	0.998	0.995	0.975	0.908	0.940	0.968	0.990	0.935

Table 6: Comparative performance analysis of LLMs with few-shot prompting strategies.

C Scale Guidebook

AI Psychological Interview Dialogue Validity Rating Scale (1-5 points)

Relevance

- 5 : AI responses are consistently closely related to user questions, emotions, and topics, highly aligned with the psychological interview context, and accurately address core user content.
- 4 : Most responses are closely related to user expressions, with only a few minor deviations that do not affect the overall communication.
- 3 : Some responses are related to user topics, but there are some deviations, although basic communication can still be maintained.
- 2 : Many responses deviate from user topics, with insufficient understanding of core user needs, and obvious obstacles in communication.
- 1 point: AI responses are almost completely detached from user topics, seriously disconnected from the psychological interview context, and frequently miss the point.

Accuracy

- 5 : Precise understanding of user questions, suggestions, and analysis provided are fully in line with psychological principles and common sense, accurately grasp and respond to user emotions.
- 4 : Basically accurately understand user questions, with slight deviations in the application of psychological knowledge and emotional judgment, but does not affect overall understanding.
- 3 : Able to understand some user questions, but there are more deviations in understanding, and the suggestions given are somewhat reasonable but not rigorous enough.
- 2 : Frequently misunderstand user questions, the information provided has many errors and may mislead users, and the emotional judgment is inaccurate.
- 1 point: Frequently and seriously misunderstand user questions, the information given completely violates psychological principles, seriously misleads users, and is unable to recognize emotions.

Completeness

- 5 : The response content is complete, well-organized and clearly structured, comprehensively covering all aspects of the user's question, with no omission of key information.
- 4 : The response content is relatively complete, including most of the key information, with slight deficiencies in details.
- 3 : Able to provide some key information, but there are many deficiencies in the overall content, and the logical structure is not clear enough.
- 2 : The response content is severely missing key information, too brief and vague, making it difficult for users to obtain effective help.
- 1 point: The response content is chaotic and provides almost no valuable information, completely failing to meet user needs.

Guidance

- 5 : Actively and proactively guides users to express themselves in depth, effectively promotes the in-depth development of the interview through appropriate methods, and has strong interactivity.
- 4 : Able to guide, the guidance method is basically natural, and it has a certain role in promoting users' in-depth expression.
- 3 : Occasionally guide users, the guidance effect is average, and sometimes it is not possible to effectively stimulate users to express themselves further.
- 2 : Rarely actively guide users, the interview is more passive, lacking interactivity and exploration.
- 1 point: Completely lacks the intention to guide users, only mechanically responds to user questions, and the interview is dull and uninteresting.

D Prompts

Depression Detection Prompt

Input

- Provide psychological counseling dialogue content.
- Determine whether the dialogue content meets the following depression diagnostic criteria and strictly follow the reasoning process.

Symptom Definitions

1. Depressed Mood

Symptom Manifestation: Subjectively feeling sad, empty, hopeless, or experiencing emotional distress; or observed crying, emotional breakdowns, or shouting.

Corresponding Questions:

- "In your life, have you ever had a period of at least two consecutive weeks where you couldn't feel happy, felt sad, empty, or hopeless every day?"
- "That period must have been tough for you. Can you tell me how that felt during those two weeks? Was there anything specific that triggered these feelings?"
- "In the past two weeks, have you felt down or empty every day?"

Scope of This Symptom:

- **Time Range:** Must last for at least two weeks.
- **Time Frequency:** Must occur "most of the day, nearly every day."
- **Symptom Reasoning Process:**
 - Both question 1 and question 3 must be answered affirmatively.
 - If question 2 is present, the described feelings must align with the symptom manifestations.
 - Only when these conditions are met, can the symptom be judged as "Yes."

2. Loss of Interest or Pleasure

Symptom Manifestation: Significant decrease in interest or pleasure in almost all activities.

Corresponding Questions:

- "Have you ever had a period of at least two consecutive weeks where you had little interest, even in things you used to enjoy?"
- "In the past two weeks, have you also been feeling uninterested and finding things meaningless?"
- "When you feel uninterested, how do you usually spend your time?"

Scope of This Symptom:

- **Time Range:** Must last for at least two weeks.
- **Time Frequency:** Must occur "most of the day, nearly every day."
- **Symptom Reasoning Process:**
 - Both question 1 and question 2 must be answered affirmatively.
 - If question 3 is present, the response should confirm a lack of interest, especially in previously enjoyable activities.
 - Only when these conditions are met, can the symptom be judged as "Yes."

3. Weight or Appetite Change

Symptom Manifestation: Significant weight loss or gain, or decreased/increased appetite, without intentional dieting.

Corresponding Questions:

- "During the two weeks of feeling depressed or uninterested, did you notice daily appetite changes or significant weight fluctuations? Did people around you comment on this?"

- "In the past two weeks, has your eating or weight changed?"
- "Have your eating habits and weight remained stable in the past two weeks? Any noticeable differences?"
- "Compared to before, has your food intake changed significantly in the past two weeks?"
- "When you felt down or uninterested, did you experience such changes daily in eating habits or weight, even though you were not intentionally dieting or overeating?"

Scope of This Symptom:

- **Time Range:** Within a past depressive episode, not limited to the recent period.
- **Symptom Reasoning Process:**
 - Both question 1 and question 5 must be answered affirmatively.
 - Questions 2, 3, and 4 should indicate unstable weight, significant appetite changes, or altered eating habits.
 - Only when these conditions are met, can the symptom be judged as "Yes."

4. Sleep Disturbance

Symptom Manifestation: Insomnia or excessive sleep.

Corresponding Questions:

- "During the two weeks when you felt depressed or lost interest, did you have trouble sleeping every night, such as difficulty falling asleep, waking up in the middle of the night, waking up too early, or sleeping too much?"
- "In the past two weeks, have you experienced the same sleep problems?"

Scope of This Symptom:

- **Time Range:** Within a past depressive episode.
- **Time Frequency:** "Sometimes" or more frequently.
- **Corresponding Question Range:**
 - "During those two weeks of feeling depressed or uninterested, did you have trouble sleeping every night, such as difficulty falling asleep, waking up in the middle of the night, waking too early, or sleeping too much?"
 - "In the past two weeks, have you had the same sleep problems?"

5. Agitation or Sluggishness in Daily Behavior

Symptom Manifestation: Physical agitation or sluggishness, or feeling mentally dull or numb.

Corresponding Questions:

- "During the two weeks of feeling depressed or uninterested, did you become sluggish or agitated almost every day? Were you noticeably restless or unable to sit still?"
- "In the past two weeks, have you had similar feelings of sluggishness or agitation?"

Scope of This Symptom:

- **Time Range:** Within a past depressive episode.
- **Time Frequency:** "Sometimes" or more frequently.
- **Symptom Reasoning Process:**
 - Both questions 1 and 2 must be answered affirmatively.
 - This condition would indicate either physical restlessness or emotional numbness.
 - Only when these conditions are met, can the symptom be judged as "Yes."

6. Fatigue or Loss of Energy

Symptom Manifestation: Persistent fatigue or a lack of energy, feeling drained.

Corresponding Questions:

- "During the two weeks of feeling depressed or uninterested, did you feel like you had no energy, almost every day?"
- "In the past two weeks, have you had similar feelings of being constantly tired or lacking energy?"

- "Was the lack of energy constant or did it occur sporadically?"

Scope of This Symptom:

- **Time Range:** Within a past depressive episode.
- **Time Frequency:** "Sometimes" or more frequently.
- **Symptom Reasoning Process:**
 - If both questions are answered affirmatively, it suggests significant lack of energy.
 - The symptoms must be noticeable almost every day.
 - Only when these conditions are met, can the symptom be judged as "Yes."

7. Feelings of Worthlessness or Excessive Guilt

Symptom Manifestation: Feeling worthless or excessively guilty, often inappropriate guilt.

Corresponding Questions:

- "During the two weeks of feeling depressed or uninterested, did you feel almost every day that you were worthless or excessively guilty?"
- "In the past two weeks, have you felt the same sense of worthlessness or guilt?"

Scope of This Symptom:

- **Time Range:** Within a past depressive episode.
- **Time Frequency:** "Sometimes" or more frequently.
- **Symptom Reasoning Process:**
 - Both questions must be answered affirmatively to indicate significant feelings of worthlessness or guilt.
 - Only when these conditions are met, can the symptom be judged as "Yes."

8. Difficulty Concentrating or Making Decisions

Symptom Manifestation: Difficulty thinking clearly or making decisions, inability to focus.

Corresponding Questions:

- "During the two weeks of feeling depressed or uninterested, did you often find it difficult to concentrate or make decisions?"
- "In the past two weeks, have you had trouble thinking clearly or making decisions?"

Scope of This Symptom:

- **Time Range:** Within a past depressive episode.
- **Time Frequency:** "Sometimes" or more frequently.
- **Symptom Reasoning Process:**
 - If both questions are answered affirmatively, it indicates difficulty concentrating or decision-making.
 - Only when these conditions are met, can the symptom be judged as "Yes."

9. Thoughts of Death or Suicide

Symptom Manifestation: Frequent thoughts of death, suicidal ideation, or actual suicide attempts.

Corresponding Questions:

- "During the two weeks of feeling depressed or uninterested, did you often think about death or have thoughts of suicide, or even plan or attempt suicide?"
- "In the past two weeks, have you had thoughts of death, suicide, or any suicide plans or attempts?"

Scope of This Symptom:

- **Time Range:** Within a past depressive episode.
- **Time Frequency:** At least once.
- **Symptom Reasoning Process:**
 - If either question is answered negatively, it would indicate the absence of thoughts of death or suicide.
 - If the answer is affirmative for either question, the symptom can be judged as "Yes."

10. Clinically Significant Distress or Impairment

Symptom Manifestation: The symptoms cause significant distress or functional impairment.

Corresponding Questions:

- "During those two weeks, did these symptoms cause significant distress or problems in your life, such as at home, work, school, social activities, or relationships?"
- "In the past two weeks, have these symptoms also caused distress or problems in your life?"

Scope of This Symptom:

- **Symptom Reasoning Process:** A score-based system determines if the impairment is significant.

Diagnosis Criteria

- **Symptoms 1-9:** At least **five** symptoms must be present, including either **depressed mood or loss of interest**.
- **Symptom 10:** Must be present.

Generalized Anxiety Disorder Detection Prompt

1. General Anxiety Across Multiple Issues

Symptom Manifestation: The individual expresses anxiety about multiple issues.

Corresponding Questions:

- "Do you feel anxious about several things in your life?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user mentions anxiety across multiple areas, mark as "Yes". If they only mention one specific concern and the response is affirmative, mark as "Yes". If no anxiety is expressed, mark as "No".
- If the answer is unclear, mark as "Uncertain".

2. Anxiety in the Past 6 Months

Symptom Manifestation: The individual has experienced excessive anxiety for more than half of the days over the past six months.

Corresponding Questions:

- "Have you been anxious for most of the past six months regarding work, school, or other activities?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user expresses significant anxiety for a period of time in the past six months, mark as "Yes". If the anxiety is recent but severe, mark as "Yes". If less than six months of anxiety is reported and mild, mark as "No".
- If the answer is unclear, mark as "Uncertain".

3. Anxiety Frequency in the Past 6 Months

Symptom Manifestation: The individual experiences frequent anxiety during the past six months, not just occasionally.

Corresponding Questions:

- "Did you experience significant anxiety more than just occasionally during the past few months?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user reports frequent anxiety over a period of time in the last six months, mark as "Yes". If the anxiety is reported as occasional or only a few times, mark as "No".
- If the answer is unclear, mark as "Uncertain".

4. Difficulty Controlling Worry

Symptom Manifestation: The individual finds it difficult to control their worry or anxiety.

Corresponding Questions:

- "Do you feel like you can't control your worries or anxiety?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user states that they have difficulty controlling their anxiety or worry, mark as "Yes".
- If the answer is unclear or ambiguous, mark as "Uncertain".

5. Restlessness or Feeling on Edge

Symptom Manifestation: The individual experiences restlessness, agitation, or tension.

Corresponding Questions:

- "Do you often feel restless, on edge, or agitated?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user reports restlessness or a sense of being on edge, mark as "Yes".
- If the answer is unrelated to this symptom, mark as "Uncertain".

6. Muscle Tension

Symptom Manifestation: The individual experiences muscle tension.

Corresponding Questions:

- "Do you experience muscle tension, such as tightness in your muscles?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user confirms muscle tension, mark as "Yes".
- If the answer is unclear or vague, mark as "Uncertain".

7. Fatigue or Exhaustion

Symptom Manifestation: The individual feels unusually tired or fatigued.

Corresponding Questions:

- "Do you feel easily fatigued or exhausted?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user expresses fatigue, mark as "Yes". If the user states that it happens occasionally, mark as "Uncertain".

8. Difficulty Concentrating

Symptom Manifestation: The individual has trouble concentrating or experiences a blank mind.

Corresponding Questions:

- "Do you find it hard to concentrate or focus, or feel like your mind is blank?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user consistently has trouble concentrating, mark as "Yes".
- If the user reports it happening occasionally, mark as "Uncertain".

9. Irritability

Symptom Manifestation: The individual experiences irritability.

Corresponding Questions:

- "Do you feel more irritable than usual?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user expresses irritability consistently, mark as "Yes".
- If irritability is occasional, mark as "Uncertain".

10. Sleep Disturbance

Symptom Manifestation: The individual experiences sleep disturbances (difficulty falling asleep, staying asleep, or non-restorative sleep).

Corresponding Questions:

- "Do you have trouble sleeping, staying asleep, or feel that your sleep isn't restful?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user reports any form of sleep disturbance, mark as "Yes".
- If the sleep issues are occasional, mark as "Uncertain".

11. Impact on Life

Symptom Manifestation: The anxiety causes significant distress or impairs the individual's functioning.

Corresponding Questions:

- "Has your anxiety caused you significant distress or affected your daily life?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user reports significant distress or impairment in their life, mark as "Yes".
- If the response is unclear or only mildly distressing, mark as "Uncertain".
- If no significant impact is reported, mark as "No".

Diagnosis Criteria

- **Symptom Criteria 1-4:** The user must meet these symptoms ("Yes" for all of these) to diagnose GAD.
- **Symptom Criteria 5-10:** At least three additional symptoms (from 5-10) must be marked "Yes" to confirm the diagnosis.

Social Anxiety Disorder Detection Prompt

1. Anxiety in Social Situations

Symptom Manifestation: The individual experiences significant or intense fear or anxiety in social situations where they might be judged by others.

Corresponding Questions:

- "Do you feel anxious because of social situations?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user expresses that anxiety occurs only occasionally, recently, or just for a few days, mark as "No". If the user indicates general, ongoing anxiety in social situations, mark as "Yes".
- If the response is ambiguous or no relevant answer is provided, mark as "Uncertain".

2. Fear of Social Judgement or Avoidance

Symptom Manifestation: The individual feels fear or anxiety in almost every social interaction, or avoids social situations due to this fear.

Corresponding Questions:

- "Does the fear almost always affect you when you're in social situations?"
- "Do you avoid social activities because of fear?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the individual experiences anxiety in almost all social situations or avoids them altogether, mark as "Yes".
- Otherwise, mark as "No".

3. Excessive Anxiety

Symptom Manifestation: The fear or anxiety is disproportionate to the actual threat or situation.

Corresponding Questions:

- "Do you feel that your anxiety or fear in social situations is more than what's actually happening?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user feels that their anxiety is excessive compared to the actual threat, mark as "Yes".
- Otherwise, mark as "No".
- If the response is unclear, mark as "Uncertain".

4. Avoidance Behavior

Symptom Manifestation: The individual avoids social interactions or seeks accompaniment due to fear or anxiety.

Corresponding Questions:

- "Do you actively avoid social situations or prefer to go with someone?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user avoids social situations or requires accompaniment frequently, mark as "Yes".
- Otherwise, mark as "No".
- If the response is unclear, mark as "Uncertain".

5. Impact on Life

Symptom Manifestation: The symptoms cause significant distress or interfere with daily functioning (e.g., work, social life, school, relationships).

Corresponding Questions:

- "Have these symptoms caused you significant distress or affected your daily life?"
- "Do these symptoms impact your ability to work, study, or have social relationships?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the user reports significant distress or impairment in their daily life due to these symptoms, mark as "Yes".
- If the user indicates no significant impact, mark as "No".
- If the response is unclear, mark as "Uncertain".

Diagnosis Criteria

- **Symptoms 1-4:** If at least one symptom is marked "No", the diagnosis is "No".
- **Symptom 5:** If "Yes" is indicated for significant distress or impairment, mark as "Yes".
- If the individual meets all the criteria and shows significant distress or impairment, mark as "Yes". Otherwise, mark as "No".

Suicide Detection Prompt

1. Suicide Ideation

Symptom Manifestation: The individual has thoughts of ending their own life, including mild or intense suicidal thoughts, thoughts about death, or expressions of hopelessness.

Corresponding Questions:

- "Have you had any thoughts about ending your life or about death in general?"
- "Do you ever feel hopeless or like life isn't worth living?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the individual reports thoughts about death or suicide, mark

as "Yes".

- If there is no mention of suicidal ideation or thoughts of death, mark as "No".
- If the response is unclear or related to mild, transient thoughts, mark as "Uncertain".

2. Suicide Plan

Symptom Manifestation: The individual has a detailed plan to carry out suicide, including specific methods, times, locations, and preparations (e.g., buying pills, writing a will).

Corresponding Questions:

- "Have you made any specific plans for how you would end your life?"
- "Do you have a method or any preparations in place for suicide?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the individual describes having a specific plan, such as a method or location, mark as "Yes".
- If there is no mention of a detailed plan or preparation, mark as "No".
- If the response is vague or there is uncertainty, mark as "Uncertain".

3. Suicide Attempt

Symptom Manifestation: The individual has taken action toward committing suicide, including behaviors such as overdosing, hanging, jumping, etc., with a high risk of lethality.

Corresponding Questions:

- "Have you ever tried to take your life or attempted suicide?"
- "Have you ever engaged in any behavior that could harm yourself severely?"

Scope of This Symptom:

- **Symptom Reasoning Process:** If the individual reports taking significant steps toward suicide, such as attempting overdose, hanging, or other lethal actions, mark as "Yes".
- If the individual denies any past suicide attempt or dangerous behavior, mark as "No".
- If the response is unclear or ambiguous, mark as "Uncertain".

Diagnosis Criteria

- If any of the three symptoms (suicide ideation, plan, or attempt) is marked as "Yes" and the severity is mild, moderate, or severe, mark the diagnosis as "Yes".
- If there is no mention of suicide-related behaviors or thoughts, or if the content is irrelevant, mark as "No".