Online ISSN: 3104-9338
Print ISSN: 3104-932X

Research on Intelligent Psychological Consultation Method Based on Dynamic Scale Embedding and Dialogue Analysis

Chenlong Xie, Ao Feng*, Zhilei Zhang

School of Computer Science, Chengdu University of Information Technology, Chengdu, China, 610225

Received: June 1, 2026

Revised: June 2, 2026

Accepted: June 5, 2026

Published online: June 11, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 7 (July 2026)

* Corresponding Author: Ao Feng (fengao@cuit.edu.cn)

Abstract. Mental health service demands in China keep rising, while traditional psychological counseling is hampered by uneven resource allocation, high costs, delayed responses and inconsistent practitioner expertise. Current intelligent counseling systems mainly adopt rule matching or basic model fine-tuning, leading to rigid dialogue, insufficient professionalism and poor capacity for active assessment and intervention. This paper presents an intelligent psychological counseling approach integrating dynamic scale embedding and dialogue analysis. Its technical framework covers dynamic scale embedding, standardized treatment plan generation, and the collaboration of fine-tuned models and Retrieval-Augmented Generation (RAG). By analyzing dialogue rhythm and semantics, the method embeds psychological scales dynamically via mutual information thresholds to realize unobtrusive assessment. Combined with user portraits and professional knowledge bases, it can automatically produce standardized treatment schemes. The joint use of domain fine-tuning and RAG also improves the empathy, guidance and professionalism of system responses. Experiments show the proposed method surpasses conventional methods in empathy, guidance, professionalism, fluency and reasoning efficiency. It supports low-cost, accessible and standardized intelligent psychological counseling, with great practical value and broad application prospects.

Keywords: *Intelligent Psychological Consultation; Dynamic Scale Embedding; Dialogue Analysis; Retrieval-augmented Generation; Standardized Treatment Plan*

1. Introduction

Currently, accelerated social pace and intensified competitive pressures have rendered public mental health issues increasingly severe. Statistics released by the Disease Prevention and Control Bureau of the National Health Commission indicate that there are 240 million individuals suffering from mental disorders across China, corresponding to an overall prevalence rate of 17.5% [1]. Nevertheless, traditional offline face-to-face psychological counseling suffers from multiple inherent limitations, including imbalanced resource allocation where high-quality counseling resources are predominantly concentrated in megacities, excessive service charges with single-session consultation fees varying from hundreds to over a thousand-yuan, complicated appointment procedures and delayed crisis intervention responses. In addition, the uneven professional competence among psychological counselors further undermines the quality and security of counseling services.

In recent years, the rapid advancement of Large Language Models (LLMs) has unlocked new prospects for intelligent psychological counseling research. Equipped with superior natural language comprehension and text generation capabilities, LLMs possess great potential to overcome the temporal, spatial and cost constraints inherent in conventional manual psychological counseling. Even so, current LLM-enabled psychological counseling systems still confront prominent limitations. Early systems built upon rule matching mechanisms and fixed dialogue scripts suffer from poor flexibility and insufficient emotional cognition capability. Models merely fine-tuned within narrow domains fail to acquire stable professional psychological knowledge, thus being highly susceptible to factual hallucinations, i.e., the generation of superficially plausible yet factually incorrect content [2]. Most existing systems mandate users to complete psychological scales at preset stages, resulting in stiff human-computer interaction, disruption of coherent dialogue flows, and a lack of capacity for proactive risk warning and dynamic psychological assessment.

To address the above limitations of existing approaches, this paper proposes and develops an intelligent psychological counseling system integrated with dynamic scale embedding, collaborative reasoning and standardized therapeutic scheme generation. The core contributions of this work are summarized as follows:

- A dynamic psychological scale embedding strategy is proposed. This strategy determines the optimal embedding moment of psychological scales via mutual information threshold constraint, enabling imperceptible and low-intrusive real-time psychological state assessment.

- A hybrid framework integrating fine-tuned LLMs and RAG-based collaborative reasoning is established. It harmonizes the empathic linguistic expression of LLMs and the precise delivery of professional psychological knowledge, thereby substantially mitigating factual hallucinations.
- A full-cycle service paradigm consisting of dialogue interaction, psychological scale evaluation, user psychological profiling and therapeutic scheme formulation is constructed. The proposed system can autonomously generate personalized and standardized intervention schemes, achieving end-to-end integration from psychological assessment to clinical intervention.
- Lightweight system deployment is realized. By adopting parameter-efficient fine-tuning and inference acceleration strategies, the proposed system effectively cuts down GPU memory occupancy and boosts the overall efficiency of model training and inference.

2. Related Work

2.1. Research Background and Current Situation

The shortage of high-quality psychological counselor resources and high service costs make it difficult for the general population to obtain continuous and professional intervention services. Meanwhile, affected by the "stigma" in social culture, many potential users are unwilling to take the initiative to seek help, which leads to the continuous aggravation of psychological problems. Intelligent psychological counseling systems can provide 7×24-hour uninterrupted services, effectively lower the usage threshold, realize early risk warning and timely intervention, and possess important social value and application prospects.

From the perspective of technological evolution, intelligent psychological counseling systems have gone through three main stages. The rule matching-based stage: the system responds through preset keywords and dialogue templates, which is poor in flexibility, single in reply mode, and incapable of handling complex emotions and open-ended questions [3]. The simple fine-tuning-based stage: researchers fine-tune pre-trained models with a small amount of psychological counseling dialogue data. Although the fluency is improved, the models lack the support of professional knowledge bases, present weak guidance ability and professional response ability, and are prone to hallucinations. The retrieval-augmented generation-based stage: RAG technology assists content generation by retrieving external knowledge bases and has been widely applied in knowledge-intensive scenarios. Nevertheless, there are still few studies that deeply combine RAG with dynamic scales and multi-turn dialogue analysis in the

field of psychological counseling, and mature integrated "assessment-intervention" solutions have not yet been formed [4].

2.2. Research Objectives and Technological Innovations

Targeting the existing limitations of current intelligent psychological counseling systems in emotional interaction, professional response generation, dynamic psychological evaluation and integrated service loop construction, this paper formulates four research objectives and correspondingly puts forward four pivotal technological innovations.

(1) Dynamic and low-intrusive embedding of psychological scales. This study abandons the conventional rigid paradigm of mandatory scale filling at fixed dialogue nodes, and establishes a dynamic triggering mechanism driven by dialogue rhythm and semantic comprehension. Psychological assessment scales can be seamlessly integrated into conversational interactions in an unobtrusive manner, enabling high-quality mental state evaluation without disrupting user interactive experience.

(2) Synergistically enhance system empathy, conversational guidance and professional response competence. This work addresses the integrated drawbacks including unnatural empathic expression, insufficient guiding ability and inaccurate professional knowledge application of prevailing models in Chinese psychological counseling scenarios. The proposed system can satisfy practical deployment requirements in both emotional comforting and rational psychological intervention.

(3) Automatic generation of standardized and executable personalized intervention schemes. This work achieves the technical advancement from general suggestive guidance to targeted precise intervention. Leveraging multi-dimensional user psychological portraits and authoritative professional knowledge repositories, the system is capable of generating clinical-compliant personalized intervention schemes tailored to individual mental characteristics.

(4) Construction of low-cost, accessible and large-scale universal psychological counseling services. By adopting parameter-efficient fine-tuning and inference acceleration techniques, the hardware resource dependency of the model is effectively reduced and system response latency is shortened. Accordingly, high-quality intelligent psychological counseling services can break the constraints of hardware costs and regional restrictions, facilitating large-scale deployment and widespread popularization.

To fulfill the aforementioned research objectives, the primary innovative contributions of this paper are summarized as follows:

(1) A novel dynamic scale embedding mechanism is proposed. A mutual information threshold-based decision engine is elaborately designed, and the Q-learning framework [5] is adopted to optimize scale embedding timing. A multi-objective reward function considering assessment completion rate, interaction disturbance level, analytical insight, user engagement and clinical practical value is further introduced. This mechanism enables precise timing selection, adaptive scale matching and unobtrusive user interaction, fundamentally resolving the conflict between traditional scale evaluation and coherent dialogue flow.

(2) A collaborative reasoning framework integrating fine-tuned language models and RAG is constructed. To compensate for the deficiencies that standalone fine-tuned models lack sufficient professional psychological knowledge and pure RAG architectures are devoid of humanistic emotional perception, this framework designs a dual-branch information processing architecture and a central fusion decision mechanism. Equipped with a dynamic weight adjustment function, it adaptively balances the output ratio of empathic dialogue content and specialized professional knowledge according to dialogue turns, psychological crisis intensity and consultation professionalism, realizing organic integration of emotional expression and domain-specific knowledge.

(3) A full-cycle closed-loop service workflow covering dialogue interaction, scale evaluation, user profiling and intervention scheme formulation is established. The proposed system integrates all functional modules ranging from natural conversational communication, real-time dynamic assessment and user psychological profile construction to automatic standardized intervention planning. It possesses comprehensive capacities including psychological perception, status evaluation, mental state judgment and targeted intervention, accomplishing the intelligent evolution from passive emotional companionship to active psychological intervention.

(4) The lightweight model optimization has been successfully implemented through an innovative three-stage progressive fine-tuning strategy, which is enhanced by a series of effective parameter optimization methods. This approach not only retains the full core performance of the original model but also achieves a remarkable reduction in GPU memory occupancy by approximately 22.4%. Additionally, it significantly cuts down training overhead by around 21.5% when compared to the classic LoRA method. Furthermore, the incorporation of optimized inference and decoding strategies has led to an impressive boost in text generation speed, reaching up to 180.92 characters per second. This advancement lays a solid technical foundation for low-cost system deployment and ensures seamless real-time human-computer

interaction, ultimately enhancing user experience and system efficiency.

3. Methodology

3.1. Overall System Architecture

The intelligent psychological counseling system proposed in this paper is designed with a hierarchical architecture, which is divided into four core functional layers as follows:

Data and Knowledge Base Layer: This layer stores desensitized real-world multi-turn conversational datasets, a psychological scale repository covering mainstream professional scales including SAS, SDS and PSS, as well as domain-specific psychological knowledge bases compiled from authoritative sources such as psychological monographs and clinical diagnostic guidelines.

Input Processing Layer: It undertakes the reception of user textual inputs and knowledge retrieval requests, and performs feature extraction and collaborative reasoning to maximize the inherent capability of the foundational model.

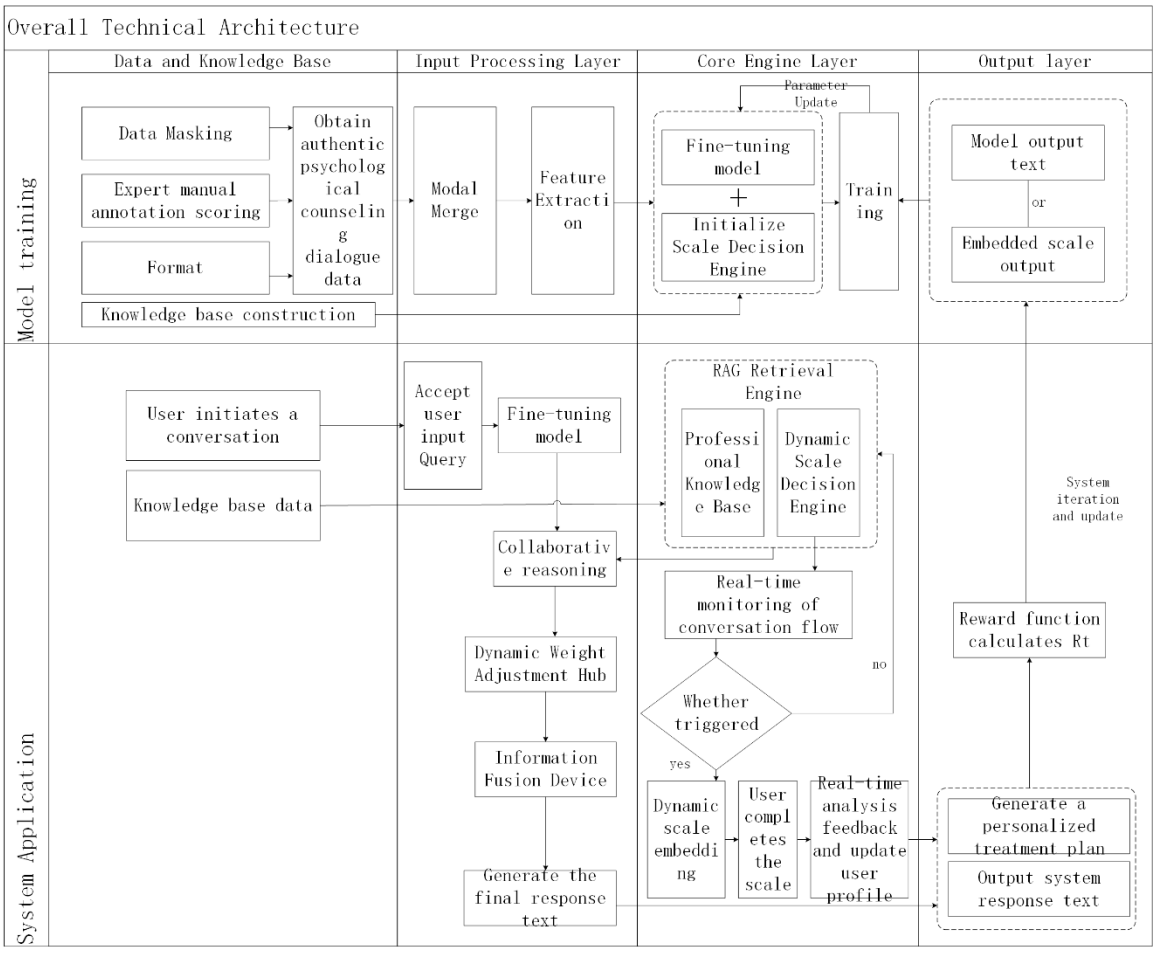

Figure 1. System Technical Architecture and Workflow Diagram.

Core Engine Layer: This layer integrates two pivotal functional engines, i.e., the scale decision engine and the RAG retrieval engine, which jointly realize dynamic judgment and analytical reasoning for scale embedding deployment [6].

Output Layer: It enables diverse human-computer interaction functions, including text-voice multimodal conversation, online scale completion, intervention scheme generation and mental health risk alert.

The overall technical framework and internal data flow of the presented system are depicted in Figure 1.

3.2. Data Collection and Expert Annotation

Given the shortage of high-quality annotated corpora in psychological counseling research, this study establishes a standardized pipeline covering data collection, filtering and manual annotation, so as to ensure the professionalism, diversity and credibility of acquired training data.

Raw conversational data are collected from authentic counseling records provided by cooperative psychological counseling institutions, covering prevalent psychological issues including anxiety, depression, interpersonal conflicts and stress regulation. All collected data are strictly processed via data desensitization in accordance with the following three core principles:

(1) Minimization Principle. Only privacy-sensitive information such as full names, contact information and specific residential addresses is desensitized, so that the semantic integrity of original dialogues can be maximally preserved.

(2) Irreversibility Principle. One-way encryption and data masking strategies are adopted to make desensitized data incapable of being restored to original versions via conventional technical means.

(3) Consistency Principle. The original dialogue format, temporal logic and textual structure remain unchanged after desensitization, which guarantees the usability of processed data for subsequent annotation tasks and model training procedures.

After acquiring desensitized raw data, domain psychological experts carry out data filtering and collation based on unified criteria. Conversational clips that conform to the logical flow of user utterance → counselor reply → user feedback → further counselor response are screened out, enabling each data sample to record the complete reasoning process from explicit emotional expression to implicit psychological demand mining. The turn number of each independent

dialogue is restricted within 3 to 8 turns, striking a good balance between sufficient contextual information and practical training efficiency.

In this work, standardized utterance templates for users and response templates for professional counselors are constructed separately. User-side templates simulate emotional expressions with distinct intensity levels. For example, mild anxiety is expressed as I have poor sleep quality recently, while severe anxiety is described as I cannot fall asleep all night, accompanied by rapid heartbeat and extreme mental breakdown. In contrast, counselor response templates adopt a unified paradigm of empathic recognition → conversational guidance → professional intervention suggestions. A typical case is illustrated as follows: I can perceive your current negative mood (empathy). How long have you been trapped in such a state (guidance)? Persistent insomnia is closely correlated with anxious emotion, and you can start with simple breathing relaxation training for relief (professional suggestion).

Combined with practical field research on mainstream counseling scenarios, the established template library covers 16 daily life scenarios including workplace pressure, family interpersonal tension and academic anxiety, together with 28 typical emotional states such as anxiety, depression, rage and loneliness. In total, 448 standardized data templates targeting representative counseling scenarios are constructed, which effectively enrich data diversity and strengthen the generalization performance of trained models.

All annotation tasks are completed collaboratively by three practitioners holding officially recognized national psychological counseling qualifications. The core annotation dimensions and corresponding five-level scoring criteria are specified as follows:

Empathy Performance: This dimension evaluates whether counselor responses contain valid empathic content and precisely match the category and intensity of users' actual emotions. Score 1 refers to absence of empathy or obvious emotional mismatch; Score 3 indicates basic empathic phrases are used yet fail to fit real emotional intensity; Score 5 represents refined scenario-based empathic expressions that perfectly align with users' inner emotional conditions.

Conversational Guidance Ability: It measures whether designed replies contain guiding inquiries and can drive dialogues to dig out deep-seated psychological needs beyond superficial emotions. Score 1 means offering direct suggestions without any guiding content; Score 3 denotes available guiding questions that cannot further deepen thematic discussion; Score 5 signifies adopting progressive questioning strategies to explore underlying demands and naturally introduce professional psychological perspectives.

Domain Professionalism: It assesses the standardization of psychological terminology usage

and compliance with formal counseling ethical norms. Score 1 stands for incorrect terminology application or violation of basic ethics such as negating users' subjective feelings; Score 3 refers to correct use of elementary professional terms without ethical conflicts; Score 5 indicates accurate and normative terminology usage as well as full ethical compliance, including respecting individual emotional cognition and avoiding compulsory intervention advice.

In the annotation phase, all qualified annotators finish independent scoring separately. For samples with score differences larger than 1 point, all three annotators conduct collective discussion to reach a consistent evaluation result. Furthermore, 10% of all annotated samples are randomly selected to compute the Kappa coefficient, ensuring that the final inter-annotator agreement coefficient is no less than 0.7.

3.3. Base Model Selection and Fine-Tuning

The choice of pre-trained base models fundamentally determines the performance ceiling of fine-tuned models and the overall deployment cost. In view of the practical characteristics of Chinese psychological counseling scenarios, this study evaluates candidate models from three core dimensions: Chinese language compatibility, GPU memory consumption and practical inference efficiency.

Targeted at Chinese conversational psychological counseling scenarios, the qualified models are required to possess powerful capabilities in Chinese semantic understanding, fine-grained emotion recognition, and perception of implicit emotional cues conveyed by colloquial modal particles. Accordingly, only pre-trained models originally optimized for Chinese language are selected as candidate alternatives.

Considering the GPU memory limitations of practical deployment platforms including local servers and cloud computing instances, this experiment adopts RTX 3090 graphics cards with 24 GB physical memory. It is necessary to strike a reasonable trade-off between model expressive capability and hardware deployment cost. Models supporting mainstream quantization strategies such as INT4 and FP8 are preferentially selected to lower practical deployment thresholds.

High interactivity is an essential requirement for psychological counseling dialogue systems, hence the end-to-end response latency must be maintained within user-acceptable ranges. In this study, we set a practical optimization target that the average generation latency of responses with less than 500 Chinese characters is controlled within 3 seconds, and further evaluate the token generation speed of candidate models under designated hardware configurations.

Based on the above practical constraints, this work conducts comprehensive preliminary assessment on three prevailing open-source Chinese large language models, namely ChatGLM3-6B [7], Baichuan2-7B [8] and Qwen-7B [9]. The experimental results demonstrate that all three models achieve satisfactory Chinese language adaptation performance. In particular, ChatGLM3-6B features lower GPU memory usage of approximately 14 GB under identical experimental conditions, satisfies the latency requirement for real-time human-computer interaction, and exhibits prominent advantages in response fluency and content safety. Consequently, ChatGLM3-6B is determined as the official base model in this research.

To empower general-purpose pre-trained models with professional psychological counseling capabilities, this paper proposes a three-stage progressive fine-tuning paradigm. The complete multi-turn user-counselor dialogue context is adopted as training corpus, and standard cross-entropy loss function is adopted for model parameter optimization. The detailed experimental configurations for each fine-tuning stage are elaborated below.

Stage 1 (Epoch 1–3): Emotional Interaction Ability Training: This stage mainly focuses on the acquisition of empathic expression and interactive guidance skills, enabling the model to master the most basic interactive logic in psychological counseling scenarios. In the designed weighted loss function, the weight coefficients of empathy and guidance supervision labels are both set to 40%, and the weight of professional knowledge supervision labels is set to 20%. The initial learning rate is configured as 3×10^{-5}, combined with warm-up learning rate scheduling strategy. After the first training stage, the average evaluation scores of empathy and guidance on the validation set can steadily reach above 3.5.

Stage 2 (Epoch 4–6): Comprehensive Capability Balanced Optimization: After the model masters basic empathic interaction and dialogue guiding logic, this stage aims to further integrate and optimize overall counseling capabilities. The loss weight of the three evaluation dimensions is adjusted to an equal ratio of 33.3%, and the learning rate is decayed to 1×10^{-5}. This optimization strategy effectively avoids the tendency that the model excessively pursues emotional resonance while neglecting normative professional replies, and realizes synchronous performance improvement across empathy, dialogue guidance and domain professionalism.

Stage 3 (Epoch 7–8): Targeted Weak Sample Reinforcement Training: Based on quantitative evaluation results on the validation set, samples with evaluation scores lower than or equal to 2 in any single dimension are screened out as low-quality deficient samples. These samples are adopted for targeted supplementary retraining with the initial learning rate of 3×10^{-5}. In this process, only partial network parameters such as top-layer transformer weights are updated,

which can effectively reinforce insufficient capabilities while effectively mitigating the risk of catastrophic forgetting of previously learned interactive rules and professional knowledge.

This study adopts a dual-mode evaluation mechanism integrating automatic quantitative scoring and expert manual review to simultaneously ensure high evaluation efficiency and reliable result accuracy. The DeepSeek-V3 [10] large language model is utilized as the automatic scoring evaluator, which rates model-generated counseling replies on the validation set with a 1–5 scoring scale from empathy, guidance and professionalism three dimensions, and calculates corresponding average scores and standard deviations. With strong comprehensive reasoning and evaluation capability verified in multiple authoritative benchmarks, DeepSeek-V3 is well-suited for large-scale batch automatic scoring tasks and greatly reduces manual evaluation workload of professional experts.

For manual review, an expert team composed of three formally qualified national psychological counselors is organized to conduct random sampling verification on automatic scoring results. Samples with evaluation deviation exceeding 1 point between machine scoring and expert subjective judgment are focused on revising, so as to guarantee the credibility of final evaluation results. The final integrated score of each evaluation dimension is calculated by the following formula.

$$S_{total} = 0.7S_{auto} + 0.3S_{manual}$$

Where S_{total} denotes the integrated comprehensive score, S_{auto} represents the score obtained via automatic model evaluation, and S_{manual} stands for the scoring result given by professional human experts.

A full-scale evaluation on the validation set is performed upon the completion of an entire three-stage fine-tuning cycle. The fine-tuning process will be terminated and the well-trained model will be preserved as the final domain-specific psychological counseling model once the average integrated scores of the three evaluation dimensions all exceed 4.0. If not, the weight allocation scheme and learning rate settings of each training stage will be optimized based on evaluation feedback, and the model will enter the next round of iterative fine-tuning.

3.4. Dynamic Scale Embedding Technique

Dynamic scale embedding serves as the fundamental technique for achieving unobtrusive psychological assessment. Its specific implementation workflow is comprehensively illustrated in Figure 2, providing a clear visual representation of the entire process involved in this innovative approach.

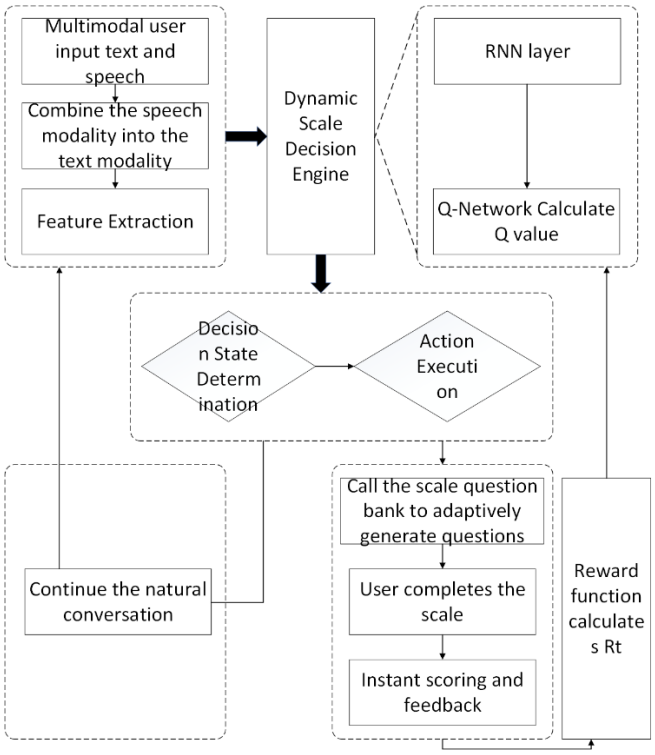

Figure 2. Dynamic Psychological Scale Embedding Framework.

The proposed framework is composed of five functional modules enclosed by dashed boxes, including feature extraction, dynamic scale decision engine, decision status discrimination, scale embedding and subsequent natural dialogue maintenance.

The detailed execution procedures are described as below: Within the dynamic scale decision engine, the Q-value represents the expected cumulative reward, which is formulated as follows:

$$Q(s,a) = E\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots \mid S_t = s, A_t = a \right]$$

$Q(s,a)$ denotes the expected cumulative discounted reward when taking action a at state s. E stands for expectation. Given the stochastic nature of responses from the environment (users), the Q-value corresponds to a probabilistic average. γ refers to the discount factor ranging from 0 to 1, which quantifies the emphasis placed on future rewards. A γ value approaching 0 indicates a myopic agent that merely prioritizes immediate rewards; conversely, a γ value close to 1 represents a farsighted agent willing to trade short-term profits for long-term returns.

At each dialogue state s_t, the system computes Q-values for all feasible actions including non-embedding, PHQ-9 embedding, GAD-7 embedding and other alternatives, and subsequently executes the action with the maximum Q-value.

The multi-objective reward function R_t evaluates the performance of action a_t, aiming to strike a balance between clinical validity and user experience. The corresponding formula is

presented below:

$$R_t = \alpha * R_{completion} + \beta * R_{perturbation} + \gamma * R_{insight} + \delta * R_{engagement} + \varepsilon * R_{clinical}$$

$R_{completion}$ denotes the completion reward, and its formula is presented below:

$$R_{completion} = I_{completed} * (1 + tanh(-\lambda * t_{abandon}))$$

Among them, $I_{completed}$ is an indicator function. If the user finishes the scale, it equals 1, otherwise it equals 0. The function of tanh is: the shorter the time $t_{abandon}$ when the user gives up completion, the greater the penalty (the lower the reward). It encourages the system to choose the time when users are more likely to finish.

$R_{perturbation}$ is the perturbation degree penalty, the formula is as follows.

$$R_{perturbation} = -\kappa * sim(s_t, s_{t-1})^{-1}$$

Herein, $sim(...)$ serves as a similarity calculation function between state s_t and s_{t-1}, such as cosine similarity. Lower similarity indicates more severe disruption of dialogue flow, which leads to a higher penalty value of $-\kappa * (...)$. This design prompts the system to insert scales in natural conversational transitions and prevent jarring breaks.

$R_{insight}$ stands for insight-related reward, and its formula is presented below.

$$R_{insight} = |score_t - E[score_{user}]|$$

The insight reward mainly calculates the absolute difference between the current scale score $score_t$ and the user's historical average score $E[score_{user}]$. A larger deviation implies that the system detects a critical clinical variation, either symptom improvement or deterioration, thereby yielding positive rewards.

$R_{engagement}$ denotes the engagement reward, and its formula is displayed below.

$$R_{engagement} = -(sentiment(s_{post}) - sentiment(s_{pre}))$$

The engagement reward quantifies the discrepancy between post-embedding utterance sentiment $sentiment(s_{post})$ and prior sentiment $sentiment(s_{pre})$. A drop in sentiment indicates degraded user experience and yields a negative discrepancy. The negative coefficient reverses this value and imposes a punitive reward. This term motivates the system to sustain users' active engagement and emotional state.

$R_{clinical}$ represents the clinical value reward, and its formula is given below.

$$R_{clinical} = \sigma(score_t - threshold)$$

Herein, σ denotes the Sigmoid function and $threshold$ refers to the clinical cutoff value of

the scale. When $score_t$ surpasses the cutoff, the term $score_t - threshold$ becomes positive, driving the Sigmoid output to surge rapidly from 0.5 to 1. Accordingly, the system can acquire considerable rewards upon identifying high-risk status. This reward component serves as the core incentive to facilitate the system's fundamental clinical efficacy.

3.5. Collaborative Reasoning of Fine-tuned Model and RAG

A collaborative reasoning mechanism integrating fine-tuned model and RAG module is proposed in this work to compensate for the limited domain knowledge of standalone fine-tuned models [11]. The corresponding workflow is illustrated in Figure 3.

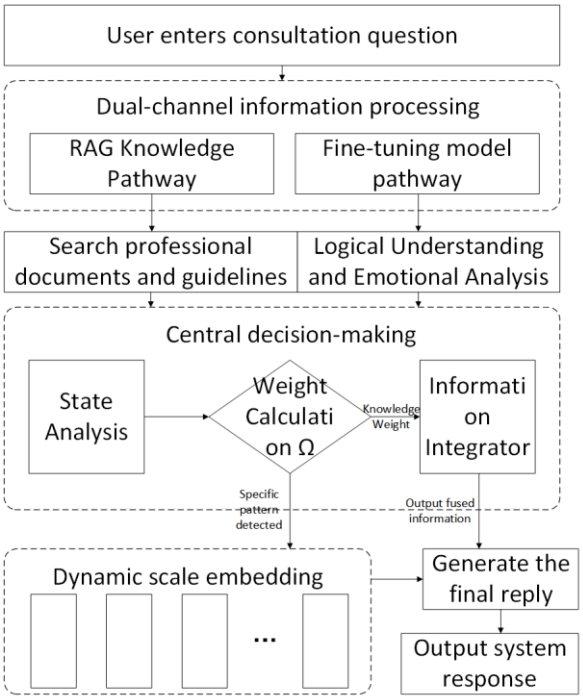

Figure 3. Framework of collaborative reasoning.

The collaborative reasoning framework mainly includes: user input consultation information, dual-path information processing, central decision-making, and system response output. The dynamic weight calculation function Ω in central decision-making analyzes the current dialogue state and determines the proportion of free generation and knowledge base reference. Its output is the knowledge weight α ranging from 0 to 1. The formula is defined as follows:

$$\alpha = \sigma(w_1 * g(t) + w_2 * S + w_3 * K - b)$$

Here, α denotes the knowledge weight that quantifies the proportion of contents retrieved from the RAG knowledge base in final responses. The Sigmoid function σ maps arbitrary values into the range (0,1) to guarantee smooth and steady weight variation. $g(t)$ is a monotonically increasing function regarding dialogue turns, driving the system to leverage

more domain knowledge as the conversation deepens. S refers to the crisis signal score derived from real-time user input analysis. Specifically, S is set to 1.0 for suicidal statements, 0.3 for insomnia-related descriptions, and 0 in the absence of sensitive keywords. K measures question professionalism via similarity calculation between user queries and domain knowledge corpus, and higher specificity and professionalism of queries yield larger K values. w_1, w_2, w_3 are manually predefined hyperparameters, which control the relative contribution of dialogue turn, crisis signal and query professionalism to weight adjustment.

3.6. Generation of Standardized Treatment Plans

Once adequate user information is collected via conversations and scale assessments, the system initiates the generation of standardized treatment schemes. The corresponding procedure is illustrated in Figure 4.

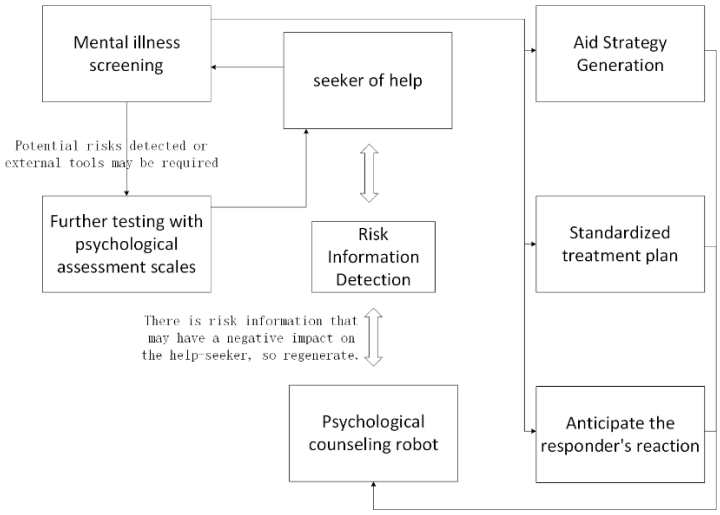

Figure 4. Interactive flowchart of standardized treatment plan.

Multimodal Sentiment Analysis and User Profile Completion. The system persistently identifies users' emotional categories (e.g., anxiety and depression) and corresponding intensities to produce systematic emotional reports. User profiles are structured and stored as knowledge graphs, encompassing static attributes (age and gender) and dynamic attributes (real-time psychological state and symptom severity). Graph-based algorithms are further applied to implement consistency validation and imputation of missing attribute values.

Multidimensional User State Mapping. Integrating scale assessment results, the system projects user status onto a multidimensional evaluation vector that covers multiple clinical dimensions, including depression severity, anxiety severity, stress level, and social function status.

Scheme Retrieval and Personalized Generation. Leveraging the aforementioned multidimensional vector, the RAG module retrieves the most clinically matched treatment schemes and intervention strategies from the domain-specific knowledge base. Subsequently, the fine-tuned model adapts the retrieved generic schemes based on comprehensive user profiles and dialogue contexts, so as to generate natural and practically feasible standardized therapeutic suggestions (e.g., "Based on your current condition, you can try the basic practice of the following cognitive behavioral therapy...").

3.7. Construction and Retrieval Configuration of Knowledge Bases

Two independent knowledge bases are established in this work, namely the scale knowledge base and professional psychological knowledge base.

In terms of scale knowledge base construction, a comprehensive database covering diverse psychological scales across multiple domains is adopted. High authority and reliability are guaranteed for the selected database to secure valid practical application. All selected scales are vectorized to convert textual content into high-dimensional vector representations, facilitating subsequent retrieval and analytical processing. A dynamic embedding-based scale retrieval system is developed to achieve efficient data query. Dynamic embedding enables real-time adjustment of retrieval strategies according to user queries, which enhances the flexibility and adaptability of the retrieval process. This technique also facilitates accurate user intention perception and improves overall retrieval accuracy.

The professional psychological knowledge base is compiled from authoritative psychological textbooks and clinical guidelines. Core contents including intervention approaches and case analysis are extracted from mainstream Chinese monographs, such as Theory and Practice of Counseling and Psychotherapy and Introduction to Clinical Psychology. Diagnostic criteria and standardized intervention workflows are sourced from official documents including Clinical Psychotherapy Guidelines issued by Chinese Mental Health Association and Guidelines for Depression Prevention and Treatment.

Contents of both databases are encoded into high-dimensional vectors by embedding models and stored in vector databases. The configured retrieval parameters are set as TopK=5 to return the top five most relevant knowledge items, with a similarity threshold of 0.7, where items scoring below the threshold are deemed irrelevant. To further optimize retrieval quality, a knowledge-dialogue fusion mechanism is designed. A dedicated knowledge filter performs secondary screening on retrieved results to remove redundant or context-conflicting knowledge snippets. For example, recommendations concerning cognitive behavioral therapy will be

discarded if the user has already received relevant treatment.

4. Experimental Results and Analysis

4.1. Experimental Settings

Experimental Environment: All experiments are implemented on a single NVIDIA RTX 3090 GPU with 24GB memory.

Baseline and Comparative Methods:

Baseline: The original unfine-tuned ChatGLM3-6B model.

Comparative Method: Model fine-tuned via Low-Rank Adaptation (LoRA).

Proposed Method: The integrated system incorporating three-stage fine-tuning, dynamic scale embedding and collaborative reasoning proposed in this work.

Evaluation Metrics:

Subjective metrics include empathy, guidance capability and professionalism. Three experienced psychological counselors perform double-blind scoring ranging from 1 to 5 on 1000 unseen test cases.

Objective metrics cover fluency, coherence and safety, which are automatically evaluated by DeepSeek-V3. Hardware and efficiency indicators involve GPU memory consumption, training duration and inference speed measured by average response latency for 500-word outputs.

4.2. Subjective Evaluation Results

Table 1 presents the subjective evaluation outcomes. The proposed method achieves substantial superiority over the baseline model on empathy, guidance and professionalism, with scores of 4.3, 4.1 and 4.5, in contrast to the baseline scores of 2.8, 2.5 and 3.1. The results demonstrate that the designed three-stage fine-tuning strategy and collaborative reasoning mechanism effectively boost the overall performance of the model for psychological counseling applications.

Table 1. Results of subjective evaluation indicators.

Evaluation Metric	Evaluation Approach	Baseline Model	Proposed Method
Empathy	Human Evaluation	2.8 ± 0.7	4.3 ± 0.5
Guidance	Human Evaluation	2.5 ± 0.8	4.1 ± 0.6
Professionalism	Human Evaluation	3.1 ± 0.9	4.5 ± 0.4

4.3. Objective Evaluation and Efficiency Comparison

The objective assessment and efficiency comparison results are illustrated in Table 2 and

Table 3. In terms of generation quality, the proposed method outperforms the baseline model with scores of 4.7 in fluency, 4.6 in coherence and 4.8 in safety.

Regarding resource consumption and operational efficiency, compared with the LoRA-based model, the GPU memory usage drops by approximately 22.4%, decreasing from 20.13 GB to 15.62 GB, and the training time is shortened by roughly 21.5%, falling from 5.26 hours to 4.13 hours. In contrast to the original baseline model, the inference speed rises by about 87.4%, increasing from 96.55 characters per second to 180.92 characters per second [12].

Table 2. Objective Quality Evaluation Results.

Evaluation Metric	Evaluation Approach	Baseline Model	Proposed Method
Fluency	Automatic Evaluation	4.5 ± 0.3	4.7 ± 0.2
Coherence	Automatic Evaluation	3.9 ± 0.5	4.6 ± 0.3
Safety	Automatic Evaluation	4.2 ± 0.6	4.8 ± 0.2

Table 3. Comparison of Training and Inference Efficiency.

Comparison Item	LoRA /Baseline Model	Proposed Method
GPU Memory Consumption	20.13 G (LoRA)	15.62 G
Training Time (20 epochs)	5.26 h (LoRA)	4.13 h
Inference Speed (500-character Output)	96.55 chars/s (Baseline)	180.92 chars/s

4.4. Functional Verification and Ablation Experiment

Table 4. Comparison of Inference Output Speed Experimental Results.

Proposed System	Baseline Model
==500-Character Output Speed Test Results ==	== 500-Character Output Speed Test Results ==
Number of test samples: 10	Number of test samples: 10
Average generated characters: 512	Average generated characters: 503
Average generated tokens: 354	Average generated tokens: 416
Average response time: 2.83 seconds	Average response time: 5.21 seconds

Functional tests based on simulated conversations prove that the system can accurately judge the timing of scale embedding. For instance, the system actively pushes the SDS scale when the user mentions insomnia and low mood for three consecutive times. After scale completion, user profiles can be updated automatically, and standardized treatment plans with concrete behavioral advice such as twice-weekly exercise recommendations will be generated without abrupt interruption of conversation flow.

Three variant models are constructed to verify the validity of individual modules: the version without dynamic scale module that relies merely on dialogue, the version removing RAG module from collaborative reasoning that only adopts fine-tuned model, and the scheme with fixed knowledge weight instead of dynamic weighting mechanism. Experimental results reveal that removing any module leads to obvious declines in professionalism and guidance scores,

with average drops of 0.9 and 0.7 points respectively, which demonstrates all modules are indispensable to the whole system.

5. Discussion

The proposed method addresses multiple defects of conventional and existing intelligent psychological counseling systems in a systematic manner. Supported by mutual information threshold and Q-learning decision framework, dynamic scale embedding integrates standardized assessment tools into natural dialogue in a non-intrusive way. It eliminates rigid forced evaluation at fixed stages, improves user acceptance and enriches collected data. Its core merit lies in autonomous timing judgment based on dialogue rhythm and semantic features, realizing the intelligent transition from passive response to active assessment.

The collaborative reasoning framework combining fine-tuned model and RAG achieves adaptive balance between empathy priority and knowledge priority via dynamic weighting mechanism. At the early dialogue stage, the system prioritizes empathetic interaction and emotional bonding with a knowledge weight of approximately 30% to build user trust. As the conversation proceeds or crisis cues are detected, the weight of RAG module rises automatically to 50% for in-depth consultation and 80% under risky circumstances, guaranteeing accurate and safe professional responses. This framework effectively reduces factual hallucinations inherent to standalone fine-tuned models, and makes up for the lack of emotional warmth in pure RAG systems [13].

The complete service loop covering dialogue, scale assessment, user profiling and treatment planning endows the system with continuous capabilities ranging from listening and evaluation to intervention. The design conforms to practical counseling procedures, where practitioners gather information, draw judgments and deliver assessments or therapeutic suggestions appropriately. Experimental results verify its rationality, with guidance score reaching 4.1 and professionalism score reaching 4.5. This approach offers a feasible technical scheme for popular and standardized mental health services free from time and cost constraints, possessing promising application prospects in communities, schools, enterprises and online platforms.

Despite satisfactory performance, this research still has limitations.

Insufficient depth of multimodal fusion: The current system only combines text and audio data, while facial expressions, physiological signals and other valuable visual and biological information for emotion recognition are not fully utilized.

Limited dataset coverage: Although common mental health scenarios are included, the

generalization ability toward specific groups such as teenagers, the elderly and patients with rare psychological disorders still needs further verification [14].

Acknowledgements

This research is supported by the Key R&D Program of Sichuan Provincial Science and Technology Plan "Research and Application of Key Technologies of Psychological Counseling Robot Based on Large Language Model" (Grant No. 2024YFFK0251). The authors gratefully acknowledge the cooperative psychological counseling institutions and relevant health care organizations for their valuable assistance in data collection, expert annotation, and practical application verification.

References

[1] Healthy China Initiative Promotion Committee. (2019). *Healthy China Initiative (2019–2030)* [Government report]. https://www.nhc.gov.cn/wsxf/zcfg/201907/2a771d89e00e4b228028335d4fcf1a7d.shtml

[2] Kermani, A., Perez-Rosas, V., & Metsis, V. (2025, May). A systematic evaluation of LLM strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. RAG. *In Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)* (pp. 172-180).

[3] Tao, D., Lee, T., Chui, H., & Luk, S. (2024, April). Modeling intrapersonal and interpersonal influences for automatic estimation of therapist empathy in counseling conversation. *In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 12692-12696). IEEE.

[4] Hu, J., Wang, A., Xie, Q., Li, Z., Ma, H., & Guo, D. (2026, March). Agentmental: An interactive multi-agent framework for explainable and adaptive mental health assessment. *In Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 31050-31058).

[5] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

[6] Dutta, A., Mruthyunjaya, S., Saddington, J., & Islam, K. S. (2025, October). Mentalic Net: Development of RAG-based conversational AI and evaluation framework for mental health support. *In 2025 IEEE International Symposium on Emerging Metaverse (ISEMV)* (pp. 77-84). IEEE.

[7] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2022, May). GLM: General language model pretraining with autoregressive blank infilling. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 320-335).

[8] Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., ... & Wu, Z. (2023). Baichuan 2: Open large-scale language models. *arXiv:2309.10305.*

[9] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... & Zhu, T. (2023). Qwen technical report. *arXiv:2309.16609*.

[10] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... & Piao, Y. (2024). DeepSeek-V3 technical report. *arXiv:2412.19437*.

[11] Kafi, M. A. A., Moni, R., & Banshal, S. K. (2026). Reasoning over recall: Evaluating the efficacy of generalist architectures vs. specialized fine-tunes in RAG-based mental health

dialogue systems. *arXiv:2601.01341.*

[12] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., ... & Stoica, I. (2023, October). Efficient memory management for large language model serving with PagedAttention. *In Proceedings of the 29th Symposium on Operating Systems Principles* (pp. 611-626).

[13] Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics, 132*, 103978.

[14] Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., ... & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, *20*(3), 318–335.