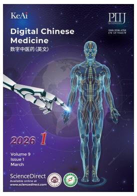
Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed



Fine-Med-Mental-T&P: a dual-track approach for high-quality instructional datasets of mental disorders in traditional Chinese medicine

Yanbai Wei[a, b], Xiaoshuo Jing[b, c], Junfeng Yan[a, b*]

a. School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

b. Hunan AI TCM Lab, Changsha, Hunan 410208, China

c. School of Traditional Chinese Medicine, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

ARTICLE INFO

Article history
Received 14 October 2025
Accepted 14 January 2026
Available online 25 March 2026

Keywords
Mental disorder
Traditional Chinese medicine (TCM)
Instruction dataset construction
Instruction tuning
Large language model

ABSTRACT

Objective To investigate methods for constructing a high-quality instructional dataset for traditional Chinese medicine (TCM) mental disorders and to validate its efficacy.

Methods We proposed the Fine-Med-Mental-T&P methodology for constructing high-quality instruction datasets in TCM mental disorders. This approach integrates theoretical knowledge and practical case studies through a dual-track strategy. (i) Theoretical track: textbooks and guidelines on TCM mental disorders were manually segmented. Initial responses were generated using DeepSeek-V3, followed by refinement by the Qwen3-32B model to align the expression with human preferences. A screening algorithm was then applied to select 16 000 high-quality instruction pairs. (ii) Practical track: starting from over 600 real clinical case seeds, diagnostic and therapeutic instruction pairs were generated using DeepSeek-V3 and subsequently screened through manual evaluation, resulting in 4 000 high-quality practice-oriented instruction pairs. The integration of both tracks yielded the Med-Mental-Instruct-T&P dataset, comprising a total of 20 000 instruction pairs. To validate the dataset's effectiveness, three experimental evaluations (both manual and automated) were conducted: (i) comparative studies to compare the performance of models fine-tuned on different datasets; (ii) benchmarking to compare against mainstream TCM-specific large language models (LLMs); (iii) data ablation study to investigate the relationship between data volume and model performance.

Results Experimental results demonstrate the superior performance of T&P-model fine-tuned on the Med-Mental-Instruct-T&P dataset. In the comparative study, the T&P-model significantly outperformed the baseline models trained solely on self-generated or purely human-curated baseline data. This superiority was evident in both automated metrics (ROUGE-L > 0.55) and expert manual evaluations (scoring above 7/10 across accuracy). In benchmark comparisons, the T&P-model also excelled against existing mainstream TCM LLMs (e.g., HuatuoGPT and ZuoyiGPT). It showed particularly strong capabilities in handling diverse clinical presentations, including challenging disorders such as insomnia and coma, showcasing its robustness and versatility. Data ablation studies showed that T&P-model performance had an overall upward trend with minor fluctuations when training data increased from 10% to 50%; beyond 50%, performance improvement slowed significantly, with metrics plateauing and approaching a saturation point.

*Corresponding author: Junfeng Yan, E-mail: junfengyan@hnucm.edu.cn.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2026.02.004

Citation: WEI YB, JING XS, YAN JF. Fine-Med-Mental-T&P: a dual-track approach for high-quality instructional datasets of mental disorders in traditional Chinese medicine. Digital Chinese Medicine, 2026, 9(1): 31-42.

Conclusion This study has successfully constructed the specialized Med-Mental-Instruct-T&P instruction dataset for TCM mental disorders proposed the systematic Fine-Med-Mental-T&P methodology for its development, effectively addressing the critical challenge of high-quality, domain-specific data scarcity in TCM, and providing essential data support for developing intelligent TCM diagnostic and therapeutic systems.

1 Introduction

As society intensifies competitive pressures, individuals face escalating mental health challenges [1]. Students race against the clock for high marks, while professionals strive relentlessly for promotion, contributing to a persistent rise in mental disorders. The deterioration in global mental health has been particularly acute. During the pandemic alone, the incidence of anxiety and depression cases accounted for more than 20% [2]. This challenge has drawn significant attention within the field of traditional Chinese medicine (TCM). Diverging from Western medicine's predominant focus on micro-level pharmacological intervention, TCM emphasizes macro-level systemic regulation through holistic pattern differentiation. Based on its principles of pattern differentiation and treatment, TCM offers personalized therapeutic strategies for mental disorders, demonstrating unique clinical advantages and potential [3]. However, the translation of this potential into scalable solutions is constrained by a critical shortage of specialized human resources. Clinical diagnosis in TCM mental disorders heavily relies on the clinical expertise of experienced physicians, creating a bottleneck in service delivery. The shortage of skilled practitioners in mental disorder diagnosis and treatment represents a fundamental challenge for healthcare systems [4].

The emergence of medical large language models (LLMs) since 2023 presents a promising avenue to address this capacity gap. Models incorporating TCM knowledge, typically developed by fine-tuning open-source general-purpose LLMs with domain-specific instructional datasets, offer a cost-effective method for rapid knowledge integration. They have demonstrated significant efficacy in medical auxiliary diagnosis and mitigating clinical workforce shortages [5]. Notable medical LLMs for general diagnosis and treatment include MING-MOE [6] and HuaTuo [7], as well as models more focused on mental health such as MindGPT [8]. Despite these advancements, existing models exhibit significant limitations in the precise diagnosis and treatment of mental disorders [9]. The fundamental reason lies in the fact that most of these models adopt construction methods from general medical models, failing to adequately adapt to the specific characteristics of mental disorders. Specifically: (i) knowledge depth deficiency: they lack systematic integration of specialized, high-quality literature from TCM mental disorders, resulting in superficial understanding of disease mechanisms and treatment principles. (ii) data quality and sensitivity challenge: mental health data is inherently sensitive and complex, requiring rigorous clinical standards for annotation and validation, which conventional data collection methods frequently fail to achieve. Consequently, these models provide only generic or surface-level response, falling short of the depth required for in-depth clinical decision support. This gap underscores the inadequacy of conventional dataset construction methodologies for the specialized domain of TCM mental health. To bridge this gap, this study employs a novel, tailored instruction dataset construction strategy to develop a high-quality, specialized dataset for TCM mental [10].

Common methods for constructing instruction datasets include manual annotation, data reconstruction, model-based generation, and human-aligned data synthesis [11]. Manual annotation can produce high-quality data, but is associated with high costs, limited diversity, and considerable subjectivity. Data reconstruction methods reduce costs and integrate cross-domain knowledge by repurposing existing structured data. However, the generated instructions may lack relevance to real-world clinical scenarios. For instance, instruction datasets comprising over 110 000 entries have been generated from open-source Chinese medicine knowledge graphs [12]. Model-based generation leverages LLMs to produce instruction data at scale, significantly enhancing efficiency. However, its quality depends on seed data and model capabilities, which can lead to the issues of low quality or redundancy. For instance, some TCM-specific models [13, 14] are trained on datasets that, despite their large volume, contain numerous grammatical errors and substandard samples, resulting in datasets that are extensive but lack precision. Human-aligned synthesis methods generate instruction-response pairs and then refine them through alignment optimization with human preferences. The quality of the final output is contingent upon the accuracy of source texts and necessitates additional annotation efforts. Practical implementations and refinements of this methodology include the MAGPIE framework by XU et al. [15], the method proposed by MA et al. [16], and the JaFIn corpus developed by TANABE et al. [17].

In conclusion, the core bottleneck of current LLMs in the field of TCM mental disorders is the lack of high-quality, specialized, and clinically deeply aligned instruction datasets. The existing construction methods are unable to ensure professional depth and data sensitivity, resulting in the models being unable to deeply understand the complex pathogenesis and syndrome differentiation and treatment principles of TCM mental disorders. Therefore, this study aims to design and implement a systematic instruction dataset construction plan for the characteristics of TCM mental disorders, to generate high-quality, specialized data that meets clinical practice needs, laying a solid data foundation for training TCM mental disorders LLMs that can provide precise and in-depth auxiliary diagnosis and treatment support.

2 Data and methods

This paper proposed the Fine-Med-Mental-T&P methodology for constructing a high-quality instruction dataset for mental disorders in TCM. The overall conceptual framework is illustrated in Figure 1. This strategy is designed to generate both theoretical and practical data, with its tactical implementation guided by a combination of human-aligned data synthesis with model-based generation techniques.

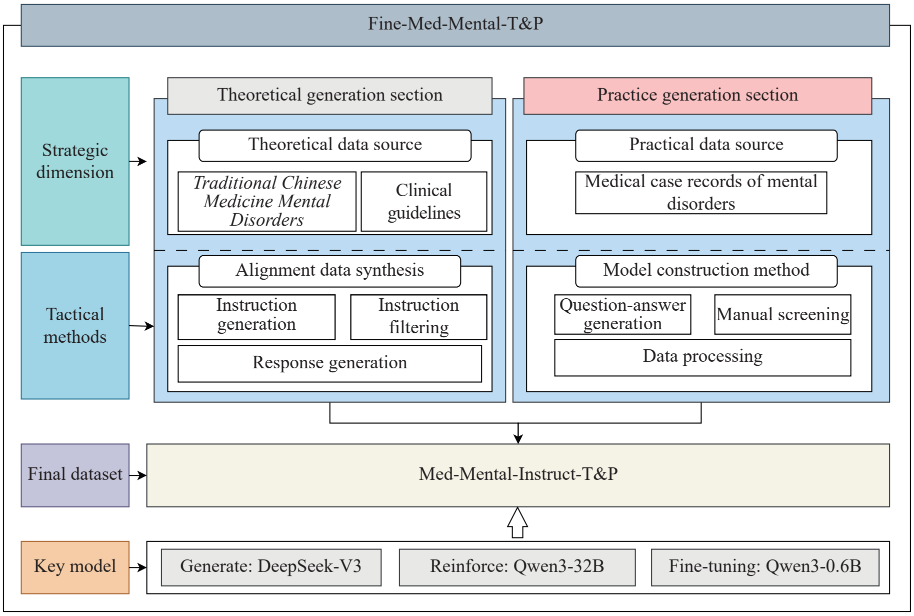


Figure 1 Overall concept of the Fine-Med-Mental-T&P methodology

2.1 Theoretical data curation

The theoretical generation process is detailed in Figure 2, which includes several key steps, such as manual segmentation, instruction screening, and alignment with human preferences.

2.1.1 Source processing and instruction generation To ground the instruction generation in established TCM theory and clinical logic, the textbook *Traditional Chinese Medicine Mental Disorders* [18] and relevant clinical guidelines [19] were selected as primary data sources.

(i) Manual segmentation and refinement. The source materials were manually segmented into coherent knowledge modules. Following principles of knowledge unit integrity, logical self-sufficiency, and appropriate length (100 – 600 words), specialists iteratively refined the segments until each represented an indivisible semantic unit. This process ensured the rigor and quality of the foundational text [20].

(ii) Prompt construction. Structured prompts were designed based on the segmented content. The prompts were crafted to be closely aligned with the core subject matter, avoid redundancy or oversimplification, and encourage diverse and in-depth responses, thereby guiding the model toward human-preferred outputs. Examples of prompts are provided in Supplementary Table S1.

(iii) Instruction generation. The constructed prompts, along with their corresponding segmented text, were input into the DeepSeek-V3 [21] model to generate the preliminary pool of instructional questions.

2.1.2 Multi-dimensional instruction scoring and filtering Following the generation of preliminary instructions, a scoring phase was implemented to assess their quality. The DeepSeek-V3 model was employed to evaluate each instruction autonomously against three defined dimensions (Supplementary Table S2). (i) Quality: evaluated the professional accuracy and theoretical fidelity of the instruction to TCM mental disorders. (ii) Complexity: assessed the inferential demand and scenario difficulty to ensure a spectrum of cognitive challenges. (iii) Relevance: measured the alignment with real-world clinical scenarios to guarantee practical utility.

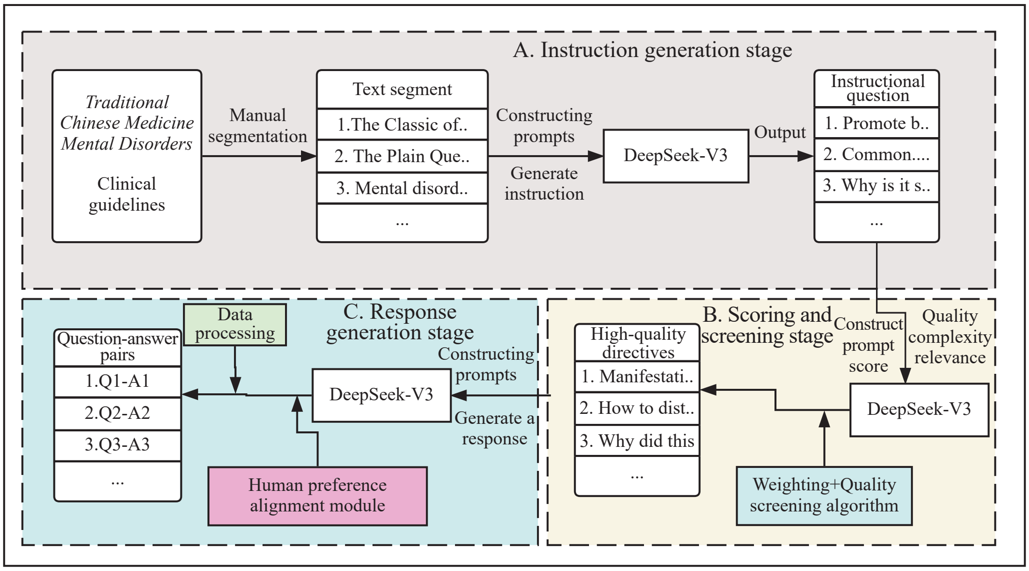


Figure 2 Workflow for theoretical knowledge-based data construction and alignment

A dual-threshold screening algorithm was constructed based on the three scoring dimensions by Equation (1) and (2). The algorithm first calculates a weighted composite score for each instruction using predefined weights: 0.4 for quality, 0.3 each for complexity and relevance. The weighting scheme was designed to prioritize content accuracy, which is the fundamental prerequisite for medical data, while assigning equal importance to training depth (complexity) and clinical applicability (relevance). Subsequently, a dual-threshold filter was applied. (i) Based on the existing literature on prompt engineering, we set the quality threshold at 7 points, where the score of 7 represents the minimum standard for clear and accurate output of LLMs. (ii) A comprehensive performance threshold of 6.5 was determined by empirical method. This threshold was optimized by analyzing the preliminary score distribution on a small validation set (n = 100) and correlating it with the experts' assessment of "acceptable" instructions. The algorithm ensures that the selected instructions meet strict quality standards while maintaining a balanced distribution across all aspects, ultimately forming a complete set of 16 495 high-quality teaching questions.

$$W_i = 0.4 \times Q_i + 0.3 \times C_i + 0.3 \times R_i \tag{1}$$

$$F_{\text{filtered}} = (W \geqslant 6.5) \cap (Q \geqslant 7) \tag{2}$$

We define four variables, namely W as the weighted composite score, Q as the quality score, C as the complexity score, and R as the relevance score.

2.1.3 Response generation and preference alignment In the response generation phase, the DeepSeek-V3 model generated multiple candidate answers for each selected instructional question. To align these responses with expert standards, the Qwen3-32B model was employed for refinement and ranking. The ranking process was guided by a structured evaluation prompt (Supplementary Table S3), which instructed the model to evaluate each candidate answer on three criteria: accuracy (alignment with TCM theory and the source text), completeness (coverage of key information points), and readability (linguistic fluency and logical clarity). Based on a composite score derived from these dimensions, the candidate answers were ranked. The highest-scoring answer for each question was selected as the final, optimized response. This methodology effectively simulates the human preference selection process for identifying optimal solutions.

2.2 Practice data expansion

The generation of practical, case-based instruction pairs followed a seed-based expansion approach, as outlined in Figure 3. Over 600 manually collected case reports on mental disorders served as seed data. These cases were formatted into structured prompts and input into the DeepSeek-V3 model for batch generation.

(i) Case collection. To ensure the comprehensiveness and scientific rigor of the dataset, clinical cases were collected from four diverse sources, with approximately 100 cases per source, with variations. The inclusion criteria for cases were complete medical records of patients diagnosed with TCM mental disorders (such as depression, insomnia, and mania), including four diagnostic methods, syndrome differentiation, treatment methods, and prescriptions. The exclusion criteria were records with severe deficiencies, ambiguous diagnoses, or incomplete follow-up information. (a) Classical TCM texts: including *Comprehensive Analysis of Verified Cases by Renowned Experts in Mental Disorders Throughout History*, *Collection of Typical Medical Cases by Contemporary Eminent TCM Practitioners*, and *Selected Verified Cases by Duzhou Liu*. (b) Modern hospital records: electronic clinical

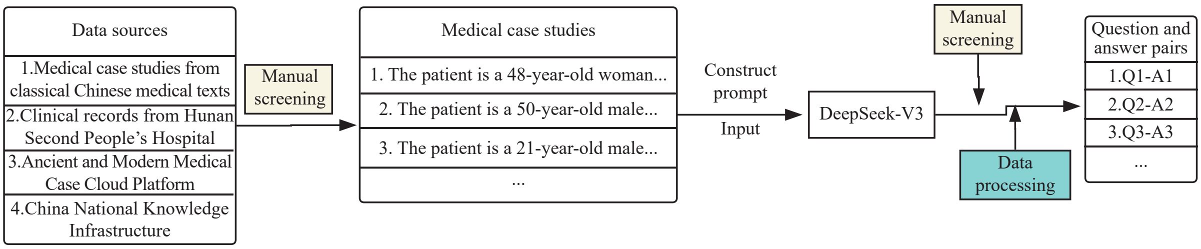

Data sources
1.Medical case studies from classical Chinese medical texts
2.Clinical records from Hunan Second People's Hospital
3.Ancient and Modern Medical Case Cloud Platform
4.China National Knowledge Infrastructure
Manual screening
Medical case studies
1. The patient is a 48-year-old woman...
2. The patient is a 50-year-old male...
3. The patient is a 21-year-old male...
...
Construct prompt
Input
DeepSeek-V3
Manual screening
Data processing
Question and answer pairs
1.Q1-A1
2.Q2-A2
3.Q3-A3
...

Figure 3 Workflow for practical (case-based) data generation

records from the TCM Diagnosis and Treatment Centre for Mental Disorders at Hunan Second People's Hospital, after strict adherence to medical data privacy and ethical norms, were extracted for outpatient and inpatient medical records that met the inclusion criteria. (c) Digital archives: case data from the "Ancient and Modern Medical Case Cloud Platform" (https://www.yiankb.com/home). (d) Contemporary academic papers and references: selected paper-reviewed case reports indexed in China National Knowledge Infrastructure (CNKI) database.

Case selection prioritized geographic representativeness (Hunan Province and surrounding regions), balanced demographic distribution (age and gender), and multi-specialty perspectives. This multi-source strategy maximized the diversity and clinical representativeness of the seed cases, which covered a broad spectrum of mental disorder conditions, presentations, and treatment regimens.

(ii) Case-based instruction generation. The collected 600 seed cases were processed into structured prompts (Supplementary Table S4) and input into the DeepSeek-V3 model for batch generation. The model was tasked with producing varied diagnostic narratives and therapeutic plans based on each prompt, thereby efficiently expanding the dataset while ensuring diversity and generalizability. This process yielded a large pool of preliminary case-based instruction pairs.

(iii) Data quality control. Throughout the data generation process, anomalous data including garbled characters and content of excessive length are first filtered out directly by algorithms. Subsequently, the generated data undergo multiple rounds of rigorous manual screening and verification by a team of TCM specialists. This ensures logical integrity and accuracy, thereby ensuring high-quality data output.

2.3 Evaluation metrics

The evaluation metric system encompassed both automated and manual categories, measuring the quality of generated text from distinct dimensions. Automated metrics, which yield scores on a scale from 0 to 1, primarily gauge formal similarity between the generated and reference texts. These metrics include BLEU-4 for overall quadruple phrase similarity, ROUGE-1 for word-level matching, ROUGE-2 for diphrase matching, and ROUGE-L for longest common subsequence matching. In contrast, human evaluation metrics employed a 0 – 10 scale, focusing on assessing the deeper quality of textual content across three dimensions, namely accuracy (evaluating the professionalism and precision of responses to mental health queries), simplicity (judging whether answers are succinct and to the point), and integrity (assessing whether responses cover all key information points).

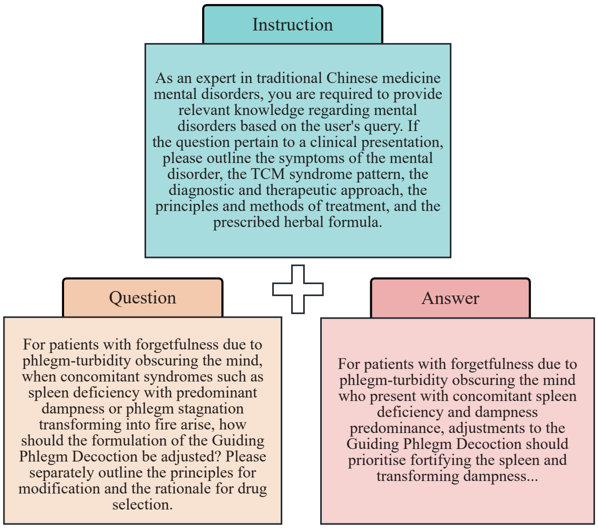
2.4 Experimental dataset

To ensure dataset quality and model application performance, three distinct datasets were constructed using different instruction generation methods to serve as fine-tuning data.

(i) Med-Mental-Self: this dataset was generated using the Self-Instruct method based on a mental disorder knowledge graph, comprising 22 000 Instruction pairs. (ii) Med-Mental-Human: this dataset was manually curated, consisting of over 1 000 high-quality instruction-response pairs for mental disorders. (iii) Med-Mental-Instruct-T&P: this dataset was constructed using the proposed Fine-Med-Mental-T&P methodology. It contains 16 000 theoretical question-answer pairs and 4 000 clinical diagnostic-therapeutic question-answer pairs for mental disorders. The dataset and associated code are scheduled to be open-sourced in February 2026 (https://github.com/hnzyy/wyb), with the aim of fostering further research and collaboration in the field of AI for TCM domain.

The test dataset employed in the study is Med-Mental-Test-9, an independent test set that was specifically developed for the purpose of conducting relevant model evaluations. It consists of 200 expert-validated cases covering nine common mental disorders (e.g., mania, insomnia, depression). Each case is formatted as a clinical question with a standardized answer. The data originate from neuropsychiatric hospital case studies and textbook question-and-answer sections.

All instruction pairs across the training and test datasets adhere to a unified structure that is composed of three distinct fields, namely instruction, input which presents the question, and output which provides the answer (Figure 4).



Instruction

As an expert in traditional Chinese medicine mental disorders, you are required to provide relevant knowledge regarding mental disorders based on the user's query. If the question pertain to a clinical presentation, please outline the symptoms of the mental disorder, the TCM syndrome pattern, the diagnostic and therapeutic approach, the principles and methods of treatment, and the prescribed herbal formula.

Question

For patients with forgetfulness due to phlegm-turbidity obscuring the mind, when concomitant syndromes such as spleen deficiency with predominant dampness or phlegm stagnation transforming into fire arise, how should the formulation of the Guiding Phlegm Decoction be adjusted? Please separately outline the principles for modification and the rationale for drug selection.

Answer

For patients with forgetfulness due to phlegm-turbidity obscuring the mind who present with concomitant spleen deficiency and dampness predominance, adjustments to the Guiding Phlegm Decoction should prioritise fortifying the spleen and transforming dampness...

Figure 4 Structure of an instruction-response pair in the Med-Mental-Instruct-T&P dataset

2.5 Experimental setup

The fine-tuning task was deployed on a parallel computing platform equipped with four NVIDIA A100 GPUs (80-GB VRAM), running on Linux #150-Ubuntu with CUDA 12.4 drivers and the PyTorch 2.60 framework. Computational efficiency was optimized using a distributed data-parallel (DDP) strategy. The Qwen3-0.6B base model was fine-tuned with low-rank adaptation (LoRA). The primary hyperparameters were set to the following values: a learning rate of 1×10^{-4}, a LoRA rank (r) of 8, a LoRA alpha (α) of 32, a LoRA dropout rate of 0.1, 30 training epochs, a per-device batch size of 4, 4 gradient accumulation steps, and a model save interval of 1 000 steps.

2.6 Experimental design for T&P-model comparison

To validate the efficacy of the proposed Fine-Med-Mental-T&P methodology, three comparative models were fine-tuned from the Qwen3-0.6B base model using the LoRA approach on different datasets: (i) T&P-model: fine-tuned on Med-Mental-Instruct-T&P dataset; (ii) Self-model: fine-tuned on the Med-Mental-Self dataset; (iii) Human-model: fine-tuned on Med-Mental-Human dataset. All models were evaluated on the independent Med-Mental-Test-9 test set. Evaluation targeted four representative mental disorders including insomnia, forgetfulness, mania, and depression. While performance was quantified via two modalities: the first being automated metrics that include bilingual evaluation understudy (BLEU)-4, recall-oriented understudy for gisting evaluation (ROUGE)-1, ROUGE-2 and ROUGE-L, and the second being manual assessment in which three TCM diagnostic experts independently rated a random sample of 30 model-generated responses for each system across three dimensions, namely accuracy, simplicity, and integrity.

2.7 Experimental design for TCM large model comparison

To further evaluate the practical utility of the Fine-Med-Mental-T&P methodology, a comparative analysis was conducted against several mainstream TCM-specialized LLMs. Two variant model were developed based on the Med-Mental-Instruct-T&P dataset: T&P_Full (full-parameter fine-tuning) and T&P_P-T (P-Tuning). These variants were evaluated alongside existing models—HuatuoGPT, ZuoyiGPT, Sunsimiao LLM, and Shuzibencao-GPT—on the Med-Mental-Test-9 dataset covering nine mental disorders. The evaluation employed the DeepSeek-V3 model as an automated judge. It scored each model's responses on a 0 – 10 scale across three dimensions—accuracy, simplicity, and integrity. Scores were recorded for each disorder, and an overall average score was calculated to assess the comprehensive competitiveness of the Fine-Med-Mental-T&P methodology.

2.8 Data ablation study design

To investigate the relationship between dataset size and model performance, a data ablation study was performed. Using the LoRA method, the Qwen3-0.6B model was fine-tuned on progressively larger subsets of the Med-Mental-Instruct-T&P dataset, including 10% (2 000 pairs), 30% (6 000 pairs), 50% (10 000 pairs), 70% (14 000 pairs) and 100% (20 000 pairs). All subsets were stratified to preserve the original theoretical-to-practical ratio (80 : 20) and the distribution of disease categories. The resulting models were evaluated on the Med-Mental-Test-9 benchmark using automated metrics (BLEU-4 and ROUGE-L). This design aimed to identify the point of diminishing returns, where additional data yields negligible performance improvement relative to the cost of data annotation.

3 Results

3.1 Performance comparison on Med-Mental-Test-9

To validate the advantages of the proposed methodology, the base Qwen3-0.6B model [22] was fine-tuned using the LoRA on three distinct datasets, which include the T&P-model that is fine-tuned on the Med-Mental-Instruct-T&P dataset, the Self-model that is fine-tuned on a self-generated dataset created by the base model itself, and the Human-model that is fine-tuned on a purely human-annotated dataset. All models were evaluated on the Med-Mental-Test-9, focusing on critical conditions, like insomnia, forgetfulness, and depression. The comparative inference performance is presented in Table 1.

The results indicated that the T&P-model achieved superior performance across all automated metrics and

Table 1 Performance comparison of the T&P-model and various models on Med-Mental-Test-9

Model	Dataset	BLEU-4	ROUGE_1	ROUGE_2	ROUGE-L	Accuracy	Simplicity	Integrity
Qwen3-0.6B	Insomnia	0.133	0.613	0.204	0.225	5.67	6.33	6.67
	Forgetfulness	0.109	0.530	0.158	0.190	5.33	5.67	3.33
	Mania	0.145	0.635	0.217	0.238	6.00	6.12	6.00
	Depression	0.141	0.619	0.218	0.227	5.67	6.00	4.67
T&P-model	Insomnia	0.546	0.816	0.626	0.637	9.00	8.67	9.00
	Forgetfulness	0.530	0.795	0.606	0.615	7.33	7.33	8.00
	Mania	0.505	0.797	0.573	0.589	8.00	8.67	8.67
	Depression	0.564	0.835	0.631	0.634	7.67	8.00	8.00
Self-model	Insomnia	0.126	0.555	0.207	0.229	7.00	6.00	8.00
	Forgetfulness	0.086	0.492	0.163	0.206	7.00	7.00	6.00
	Mania	0.097	0.534	0.177	0.208	6.67	6.00	7.00
	Depression	0.116	0.548	0.187	0.221	6.00	6.67	5.33
Human-model	Insomnia	0.130	0.606	0.205	0.221	4.67	4.00	4.87
	Forgetfulness	0.101	0.514	0.158	0.191	6.00	6.00	6.00
	Mania	0.159	0.640	0.228	0.242	5.67	6.00	6.00
	Depression	0.157	0.627	0.226	0.237	6.00	5.00	6.00

human assessment dimensions. In automated evaluation, its BLEU-4 score consistently exceeded 0.5, and its ROUGE-L score remained above 0.55 across tests (Figure 5). In manual expert evaluation, the T&P-model achieved scores of 7 or above in terms of accuracy (with the score ranging from 0 to 10), and it outperformed other models in terms of simplicity and integrity.

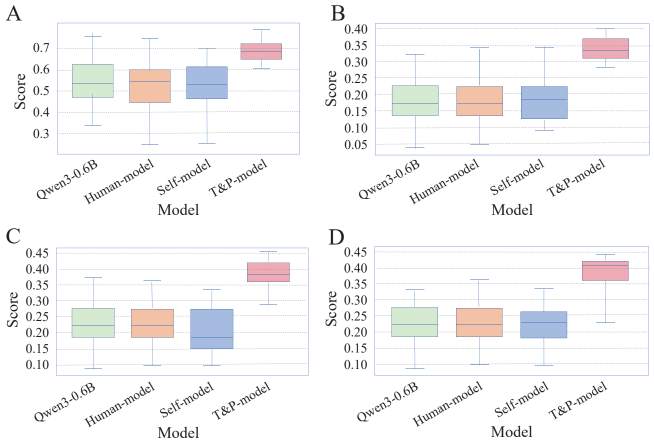


Figure 5 Comparison of ROUGE-L scores of the four fine-tuned models on different datasets of mental disorders

A, insomnia dataset. B, forgetfulness dataset. C, mania dataset. D, depression dataset.

3.2 Comparison with existing TCM LLMs

To further assess the Fine-Med-Mental-T&P methodology, two fine-tuning strategies—full parameter fine-tuning and P-Tuning—were applied to the Med-Mental-Instruct-T&P dataset, resulting in the T&P_Full and T&P_P-T models, respectively. These models, along with existing TCM LLMs (e.g., HuatuoGPT, ZuoyiGPT), were evaluated on the Med-Mental-Test-9 dataset. The DeepSeek-V3 model served as an automated judge, scoring responses across nine mental disorders.

As shown in Table 2, the T&P_Full model exhibits stable and superior performance across most disorders. It achieved notably high scores (8.22 – 8.85) for more complex conditions such as dementia, coma, forgetfulness, and vertigo, indicating strong diagnostic precision. However, its performance was slightly lower for mania (8.22) and depression (8.67), showing a discernible gap compared to the ShuzibencaoGPT model. This discrepancy may be attributed to a relative data imbalance within a training set, resulting in less robust generalization for

Table 2 Evaluation scores for TCM LLMs

Model	Insomnia	Dementia	Coma	Forgetfulness	Mania	Vertigo	Depression	Restless	Stroke
HuatuoGPT	8.67	8.37	8.67	8.67	8.44	8.44	8.56	7.89	8.44
ZuoyiGPT	8.67	8.47	8.89	8.67	8.67	8.67	8.33	8.67	7.44
Sunsimiao LLM	7.11	8.33	7.22	6.00	7.44	8.00	7.78	7.56	7.78
ShuzibencaoGPT	9.22	8.33	8.67	8.33	9.22	7.00	9.33	9.11	9.33
T&P-model	8.67	8.67	8.91	8.89	8.22	8.93	8.37	8.67	8.67
T&P_Full	8.33	8.64	8.85	8.64	8.22	8.67	8.67	8.67	8.64
T&P_P-T	6.67	6.67	8.11	7.56	6.87	7.89	6.87	6.33	8.00

these specific conditions. In contrast, the T&P_P-T model underperformed. The effectiveness of P-Tuning is highly dependent on the design of its continuous prompt embeddings. Suboptimal prompt design or inadequate task-specific tuning for certain TCM symptom patterns may have limited the model's adaptability and comprehension during evaluation.

3.3 Data efficiency results

Figure 6 presents the performance of the T&P-model fine-tuned on progressively larger fractions of the Med-Mental-Instruct-T&P dataset, averaged across the four evaluated conditions. The results revealed an overall upward trend in model performance as the training data increased from 10% to 50%. This phase indicates that the model efficiently assimilated core patterns and task-relevant knowledge despite minor fluctuations, establishing a strong performance baseline with the first half of the dataset.

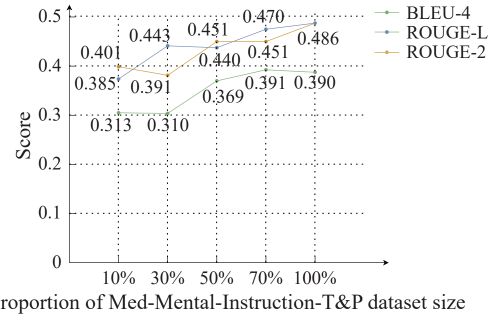


Figure 6 Performance metrics across different dataset sizes

Beyond the 50% data threshold, a trend of slowed performance improvement was observed rather than a clear decline. The performance curves began to plateau, most notably between the 70% and 100% data points. These findings suggest that for the Qwen3-0.6B model under the LoRA fine-tuning paradigm, simply augmenting the dataset beyond a certain volume yields progressively smaller performance benefits, indicating the approach of a performance saturation point.

3.4 Dataset analysis

Following the described methodology, a high-quality instruction dataset comprising 20 000 pairs was constructed. This section analyzes the dataset' scale, structure, and intrinsic quality.

3.4.1 Scale and structural analysis The Med-Mental-Instruct-T&P dataset contains 20 000 high-quality instruction pairs, adhering to a dual-source strategy of theory and practice. Theoretical instruction pairs, a set of 16 000 pairs making up 80% of the dataset, encompass nine core syndromes defined in *Traditional Chinese Medicine Mental Disorders*, namely insomnia, dementia, coma, forgetfulness, mania, vertigo, depression, restless, and stroke (Figure 7).

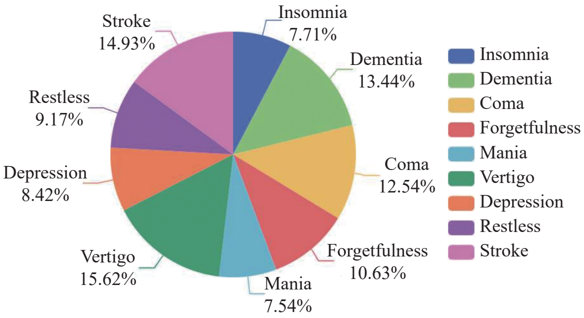


Figure 7 Distribution of responses for nine mental disorders

Among these, vertigo-related queries were the most frequent (15.62% of theoretical pairs), while manic episodes were the least frequent (7.54%). Overall, the distribution across disease categories was balanced, ensuring comprehensive coverage.

Practical instruction pairs, consisting of 4 000 pairs that represent 20% of the overall dataset, are derived from 600 clinical seed cases and primarily take the form of diagnostic and therapeutic question-answer pairs constructed based on real-world clinical scenarios.

3.4.2 Intrinsic quality assessment The intrinsic quality of the dataset was evaluated using the DeepSeek-V3 model. As shown in Figure 8, 69.95% of all instruction pairs achieved the highest quality score of 9, confirming the dataset's overall high standard.

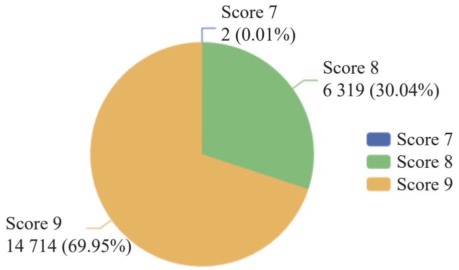


Figure 8 Quality score distribution of the Med-Mental-Instruct-T&P dataset

3.4.3 Text complexity analysis To evaluate the linguistic complexity of the Med-Mental-Instruct-T&P dataset, we conducted statistical analysis of text lengths across instruction pairs. The descriptive statistics for question and answer lengths are summarized in Table 3.

The Med-Mental-Instruct-T&P dataset exhibits a balanced text complexity profile. Questions average 42 words (median is 37.00), with 75% under 46 words, indicating concise clinical queries. Answers are substantially more detailed, averaging 176 words (median is 171.00), reflecting comprehensive explanations.

Table 3 Text length statistics of instruction pairs (characters)

Metric	Question_length	Answer_length
Mean	42.07	175.85
Standard deviation	25.06	48.74
Minimum	13.00	22.00
Median	37.00	171.00
75th percentile	46.00	201.00
Maximum	598.00	713.00

4 Discussion

4.1 Research contributions and their implications

To address the dual challenges of low-precision, large-scale instruction datasets, and shortage of diagnostic resources in mental disorders, this study proposed the Fine-Med-Mental-T&P methodology for constructing high-quality, domain-specific instruction datasets for TCM mental disorders. This strategy synergistically integrates theoretical knowledge from TCM mental disorders and clinical guidelines with practical evidence from real-world diagnostic cases. Leveraging the domestically developed DeepSeek-V3 for generation and implementing a rigorous human-in-the-loop quality assurance protocol, the methodology produced the high-quality Med-Mental-Instruct-T&P dataset containing 20 000 pairs of instructions, which comprehensively covers the theoretical knowledge and diagnostic methods of key TCM mental disorders such as insomnia, dementia, mania, and depression. This work establishes a robust data foundation for advancing AI application in TCM mental health [23].

To further validate the effectiveness of this dataset and methodology, we established multiple benchmark datasets for rigorous evaluation, including the training datasets Med-Mental-Self, Med-Mental-Human, and the test set Med-Mental-Test-9 covering nine common mental disorders. Through a multi-dimensional evaluation framework that combines expert assessment and automatic scoring by large models, we conducted comprehensive experimental verification. The results consistently demonstrated that the Med-Mental-Instruct-T&P dataset has significant advantages in quality and performance, fully proving that the Fine-Med-Mental-T&P methodology we proposed is reliable and efficient. This work ultimately laid a solid data foundation for promoting the in-depth application of artificial intelligence in the field of TCM mental health.

4.2 Analysis of evaluation metric applicability

To enhance the comprehensiveness and rigor of model evaluation, this study critically analyzed the applicability and limitations of conventional automated metrics—specifically BLEU-4 [24] and ROUGE-L—in the context of TCM mental disorder question-answering tasks.

While BLEU-4 and ROUGE-L are standard metrics for evaluating surface-level text similarity in general natural language processing (NLP), their utility for evaluating medical Q&A systems, particularly within TCM, is limited for two principal reasons.

(i) Inadequacy for semantic and logical assessment (BLEU-4). BLEU-4 based on n-gram overlap, is well-suited for tasks like translation where lexical match is paramount. However, it fails to capture the semantic comprehension and the symptom-pathogenesis logic essential to TCM. Given the complexity of TCM syndrome differentiation, relying on phrase overlap cannot accurately gauge a model's capability in theoretical adaptation and diagnostic reasoning, potentially underestimating its practical medical utility.

(ii) Overemphasis on structure over theoretical depth (ROUGE-L). ROUGE-L assesses similarity via the longest common subsequence (LCS), reflecting fluency and structural coherence. In the rich and nuanced domain of TCM, this can over-prioritize textual consistency while overlooking theoretical rationale, logical deduction, and clinical applicability of the knowledge presented [25]. TCM diagnosis hinges not on textual resemblance, but on the accurate deduction of pathogenesis from specific symptoms and the proposal of coherent treatment plans [26].

Therefore, although instrumental for general purposes, BLEU-4 and ROUGE-L exhibit clear limitations for domain-specific evaluation in TCM. The manual evaluation dimensions introduced in this study—accuracy, simplicity, and integrity—address this gap by emphasizing the logical coherence and diagnostic depth of the model's reasoning process from symptoms to treatment [27]. This human-centric framework more faithfully reflects the practical diagnostic capability and theoretical alignment of TCM question-answering systems [28].

Future work should focus on developing specialized, domain-aware evaluation metrics for TCM. Incorporating measures of semantic consistency with TCM pattern differentiation principles could significantly improve the accuracy and practical relevance of model assessment.

4.3 Methodological innovation and comparative advantage

The proposed Fine-Med-Mental-T&P methodology enhances the quality of TCM mental disorder instruction data through two synergistic innovations: (i) the dual-source integration of theoretical knowledge (textbooks/guidelines) and practical evidence (authentic clinical cases), and (ii) a dual-track quality control mechanism combining algorithmic screening with expert verification.

The core strengths of this methodology lie in its rigorous quality control and its effective integration of theory with practice. First, it grounds the dataset in authoritative theoretical sources (*Traditional Chinese Medicine Mental Disorders*, clinical guidelines) [18, 19] while enriching it with 600 real-world clinical cases. This ensures the generated instructions possess both professional depth and practical relevance. Second, it implements a robust, multi-stage quality assurance protocol. During the theoretical data generation, knowledge units were manually segmented and reviewed, followed by algorithmic filtering based on multi-dimensional scores (quality, complexity, and relevance). This human-machine collaborative loop ensures high fidelity from source curation to final output.

From a technical implementation perspective, the study consistently employed domestically developed LLMs (DeepSeek-V3 for generation, Qwen3-32B for alignment). This choice aligns with practical considerations for data handling and contributes to the application and validation of the domestic AI technology stack in specialized medical domains. These design choices not only address the long-standing challenge of insufficient high-quality instruction data in TCM mental disorder research but also establish a replicable framework for dataset construction in other specialized TCM fields.

5 Limitations and future outlook

While the Fine-Med-Mental-T&P methodology and the resulting Med-Mental-Instruct-T&P dataset have achieved significant advancements, several limitations warrant attention, and outline clear directions for future research [29]. First, the current quality control process heavily relies on experts, which limits efficiency and scale. The future focus will shift to developing automated and semi-automated technologies, such as using fine-tuned models for semantic segmentation and developing intelligent screening algorithms based on knowledge graphs, to build an efficient human-machine collaboration process. Second, the existing evaluation metrics cannot capture the deep semantics and clinical logic specific to traditional Chinese mental disorders, and manual evaluation is highly subjective. Therefore, in the future, a domain-specific evaluation system needs to be constructed. By defining core dimensions, developing semantic tools based on ontologies, and integrating automatic scoring and expert calibration composite models, more accurate and scalable evaluations can be achieved. Finally, the single data source and insufficient generalization ability are also significant challenges. The future direction lies in systematically expanding data sources, including extracting knowledge from classic medical texts, conducting multi-center clinical collaboration to collect diverse cases, and exploring generative models to enhance rare disease data, ultimately improving the generalization ability and clinical applicability of the model.

6 Conclusion

This study successfully developed the Med-Mental-Instruct-T&P, a high-quality instruction dataset resource for TCM mental disorders, and proposed the Fine-Med-Mental-T&P systematic construction methodology. This methodology innovatively integrates theoretical knowledge with clinical practice through a dual-track strategy, supported by a rigorous human-in-the-loop quality control process, thereby addressing the critical bottleneck of scarce, high-quality data in this specialized domain. Experimental results demonstrate that models fine-tuned on this dataset achieved significant improvements in both domain-specific accuracy and practical utility. This work establishes a robust data foundation for developing precise and reliable intelligent diagnostic systems in TCM mental disorders and offers a replicable technical paradigm for constructing high-quality datasets in other verticals of TCM.

Fundings

Key Scientific Research Project of the Hunan Provincial Department of Education (23A312).

Ethical statement

The data of this study have been approved by the Hospital Medical Ethics Committee of Hunan Second People's Hospital (Approval No. 2023K018).

Author contributions

Yanbai Wei: conceptualization, methodology, data curation, software, formal analysis, investigation, writing – original draft, and visualization. Xiaosuo Jing: methodology and data curation. Junfeng Yan: conceptualization, resources, supervision, project administration, and funding acquisition. All authors approved the submission and take responsibility for this manuscript.

Competing interests

Junfeng Yan is an editorial board member for *Digital Chinese Medicine* and was not involved in the journal's review or the decision related to publish this article. All authors declared that there are no competing interests.

References

[1] LIAO Y, JING X, YAN J, et al. Theoretical exploration of pathogenic factors in mental disorders based on the Huangdi Neijing. Chinese Ethnic and Folk Medicine, 2025, 34(8): 14–17.

[2] LV W, GAO M, ZHANG G. Research progress on public sleep and mental health after the COVID-19 pandemic. Psychological Monthly, 2024, 19(11): 215–217.

[3] JIANG X, ZHANG S. Discussion on the pathogenesis characteristics and treatment methods of depression in TCM. Chinese and Foreign Medical Research, 2024, 3(25): 157–159.

[4] ZHOU C, ZHENG L, LI Y. Preliminary exploration of talent development approaches in TCM psychopathology. Journal of Lishizhen Traditional Chinese Medicine, 2020, 31(9): 2251–2253.

[5] GE H. Multiple large models emerge in TCM field: the "AI old TCM practitioner" arrives. Yicai Global, 2024-12-05.

[6] LIAO Y, JIANG S, WANG Y, et al. MING-MOE: enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts. bioRxiv, 2024. doi: 10.1101/2404.09027.

[7] WANG HC, ZHAO SD, QIANG ZW, et al. Knowledge-tuning large language models with structured medical knowledge bases for trustworthy response generation in Chinese. ACM Transactions on Knowledge Discovery from Data, 2025, 19(2): 1–17.

[8] ZHANG S, ALAM E, BABER J, et al. MindGPT: advancing human-AI interaction with non-invasive fNIRS-based imagined speech decoding. bioRxiv, 2024. doi: 10.1101/2408.05361.

[9] NETEASE. Milestone in digital intelligence for TCM: Daijing TCM releases "Qihuang asking questions · large model". 163.com, 2023-07-28. Available from: https://www.163.com/dy/article/IAOP574U0514E3P4.html.

[10] JI X, ZAN H, CUI T, et al. Review of Chinese medical large language models: progress, evaluation and challenges. Chinese Journal of Information Science, 2024, 38(11): 1–12.

[11] ZHU M, SHA J, FENG C. Construction of a Tibetan instruction dataset for large language models. Chinese Journal of Information Science, 2024, 38(12): 83–96.

[12] MICHAEL-WZHU. ShenNong-TCM-LLM: repository for ShenNong-TCM-LLM (the "ShenNong" large language model, the first Chinese large language model specialised in TCM). GitHub, 2023-06-25. Available from: https://github.com/michael-wzhu/ShenNong-TCM-LLM.

[13] ZHANG J, YANG S, LIU J, et al. Empowering the revitalisation of traditional Chinese medical texts through AIGC: the construction of the Huang-Di large language model. Library Forum, 2024, 44(10): 103–112.

[14] YU H, CHENG T, CHENG Y, et al. FineMedLM-o1: enhancing the medical reasoning ability of LLM from supervised fine-tuning to test-time training. bioRxiv, 2025. doi: 10.1101/2501.09213.

[15] XU Z, JIANG F, NIU L, et al. Magpie: alignment data synthesis from scratch by prompting aligned LLMs with nothing. bioRxiv, 2024. doi: 10.1101/2406.08464.

[16] MA Y, MIZUKI S, FUJII K, et al. Building instruction-tuning datasets from human-written instructions with open-weight large language models. bioRxiv, 2025. doi: 10.1101/2503.23714.

[17] TANABE K, SUZUKI M, SAKAJI H, et al. JaFIn: Japanese financial instruction dataset. 2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr) 2024: 1-10. doi: 10.1109/cifer62890.2024.10772973.

[18] ZHAO YH, CAI DF. Traditional Chinese Medicine Mental Disorders. Shanghai: Shanghai University of Traditional Chinese Medicine Press, 1970.

[19] ZHAO Y. Clinical diagnosis and treatment guidelines for mental disorders in TCM. Harbin: Heilongjiang Mental Health Hospital, 2016.

[20] LI M, LUO X, ZHU B. Data mining of TCM formulas from classical texts and construction of a knowledge question-answering system. Library Forum, 2025, 45(4): 49-59.

[21] LIU A, FENG B, XUE B, et al. DeepSeek-V3 technical report. bioRxiv, 2024. doi: 10.1101/2412.19437.

[22] YANG A, LI A, YANG B, et al. Qwen3 technical report. bioRxiv, 2025. doi: 10.1101/2505.09388.

[23] GAO Y, ZHANG Q, ZHU S. Research on the construction of a multimodal RAG intelligent Q&A system for ancient agricultural books integrating LoRA and Qwen3-VL. Library Journal, 2025: 1-15. Available from: https://link.cnki.net/urlid/31.1108.g2.20260109.1235.010.

[24] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2001: 311-318.

[25] LIN CY. Rouge: a package for automatic evaluation of summaries. Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004: 74-81.

[26] ZHANG Q, CHEN PF, FENG LK, et al. PeMeBench: a benchmark method for Chinese pediatric medical Q&A. Big Data Research, 2024, 10(5): 28–44.

[27] WANG XY, YANG T, SUN XH, et al. Construction and application of a large model for TCM syndrome differentiation and treatment based on knowledge distillation. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2026(1): 296–304.

[28] CAO L, XU L, ZHANG YJ, et al. Standardized evaluation of large language models in the field of traditional Chinese medicine. Journal of Nanjing University of Traditional Chinese Medicine, 2024, 40(12): 1383–1392.

[29] ZHANG S, WANG W, PI XT, et al. Advances in the application of traditional Chinese medicine using artificial intelligence: a review. The American Journal of Chinese Medicine, 2023, 51(5): 1067–1083.

(Editor-in-Charge Jie Deng)

Fine-Med-Mental-T&P：一种构建高质量中医神志病指令数据集的双轨方法

魏彦柏[a, b], 荆晓朔[b, c], 晏峻峰[a, b*]

a. 湖南中医药大学信息学院, 湖南 长沙 410208, 中国
b. 湖南人工智能中医实验室, 湖南 长沙 410208, 中国
c. 湖南中医药大学中医学院, 湖南 长沙 410208, 中国

【**摘要**】**目的** 探究构建中医神志病高质量指令数据集的方法，并验证其有效性。**方法** 我们提出了 Fine-Med-Mental-T&P 这一方法论，用于构建中医神志病学领域的高质量指令数据集。该方法通过双轨策略将理论知识与实际案例研究相结合。（1）理论轨道：对中医神志病的教材和指南进行了人工分割。使用 DeepSeek-V3 生成初始响应，然后通过 Qwen3-32B 模型进行优化，以使表达符合人类偏好。随后应用筛选算法，筛选出 16 000 对高质量的指令对。（2）实践轨道：从超过 600 个真实的临床病例种子开始，使用 DeepSeek-V3 生成诊断和治疗的指令对，随后通过人工评估进行筛选，最终得到 4 000 对高质量的实践导向的指令对。两个轨道的整合形成了 Med-Mental-Instruct-T&P 数据集，总共有 20 000 对指令。为了验证数据集的有效性，我们进行了 3 项实验评估（包括手动评估和自动化评估）：（1）对比研究，以比较在不同数据集上微调后的模型的性能；（2）基准测试，与主流中医大语言模型进行比较；（3）数据消融研究，以探究数据量与模型性能之间的关系。**结果** 实验结果表明，基于 Med-Mental-Instruct-T&P 数据集进行微调的 T&P-model 具有卓越的性能。在对比研究中，T&P-model 明显优于仅基于自动生成或纯粹人工筛选的基准数据训练的基线模型。这种优越性在自动指标（ROUGE-L > 0.55）和专家人工评估（准确率评分超过 7）中均有所体现。在基准比较中，T&P-model 也优于现有的主流中医大型语言模型（例如 HuatuoGPT 和 Zuoyi-GPT）。它在处理各种临床表现方面表现出特别强大的能力，包括诸如失眠和昏迷等具有挑战性的病症，展示了强大的全面综合竞争力。数据消融研究显示，当训练数据从 10% 增加到 50% 时，T&P-model 的性能呈现出总体上升的趋势，虽然存在一些小幅波动；超过 50% 后，性能提升的速度显著放缓，各项指标趋于稳定并接近饱和点。**结论** 本研究成功构建了针对中医神志病的专业化 Med-Mental-Instruct-T&P 指令数据集，并提出了 Fine-Med-Mental-T&P 方法，有效地解决了中医领域中高质量、特定领域数据稀缺这一关键难题，为开发智能中医诊断和治疗系统提供了必要的数据支持。

【**关键词**】神志病；中医；指令数据集构建；指令微调；大语言模型