GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

International Journal of Engineering Advances

Article

Application of Large Language Model in Mental Health Clinical Decision Support System

Danyating Shen [1,*]

[1] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[*] Correspondence: Danyating Shen, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Abstract: With the rapid advancement of artificial intelligence, large language models (LLMs) have emerged as potentially transformative auxiliary tools for clinical judgment and decision-making within the mental health domain. This paper explores systematic methodologies for integrating these models into psychiatric practice, focusing specifically on technical applications in semantic modeling, sophisticated data processing strategies, and the refinement of prompt engineering alongside output optimization. A robust architectural framework and integration scheme tailored to the specificities of mental health services are constructed, providing a structured approach to clinical deployment. Furthermore, the study establishes comprehensive evaluation model standards and rigorous risk control measures designed to enhance the functional stability and ethical safety of these systems. By addressing critical concerns such as interpretability and clinical reliability, this research aims to provide a clear technical trajectory and implementation support for achieving intelligent, evidence-based, and highly understandable auxiliary decision-making in modern mental health care environments.

Keywords: large language model; mental health; clinical decision support; prompt engineering; interpretability

Received: 08 November 2025
Revised: 27 December 2025
Accepted: 10 January 2026
Published: 14 January 2026

1. Introduction

At present, mental health problems are becoming increasingly serious. The application of clinical decision support systems is conducive to improving treatment efficiency and accuracy. In recent years, based on their excellent natural language understanding and generation capabilities, large language models in the field of artificial intelligence have become an important technology to accelerate the development of the new generation of intelligent clinical auxiliary decision-making systems. They have shown good application prospects in dimensions such as processing text information of poorly structured cases, simulating doctor-patient dialogue and communication, assisting in diagnosis and providing treatment opinions [1]. This study focuses on the role of large language models in mental illness management, elaborating on their technical core principles, system framework, assessment methods and potential risk response strategies, with the aim of providing a reference for the theoretical framework and practice of intelligent mental health management of this system.

2. Theoretical Basis and Technical Principles

2.1. Overview of the Mental Clinical Decision Support System

The Clinical Decision Support System (CDSS) for mental health refers to a system that uses intelligent software based on combined medical information databases and patient information to assist psychiatrists in making choices in treatment plan decisions. Compared with the traditional CDSS, the mental health system needs to deal with issues such as higher-dimensional disordered texts (such as consultation session records, patients' self-reports, etc.), the basis for ambiguous diagnosis, and multi-source data. The current solutions that adopt preset rule sets or traditional machine intelligence algorithms are difficult to capture deep semantic relationships, and their scalability and interactivity are also poor. Therefore, it becomes particularly necessary to develop new natural language CDSS that can understand and create context patterns [2].

2.2. Structure and Principles of Large Language Models

Large language models (LLMS) take Transformer as the basic framework. They establish the semantic context of the text through multiple layers of self-attention layers and feedforward networks. The multi-layer Transformer gradually extracts the semantics of the input sequence that has been embedded and positional encoded. Its core mechanism is self-attention, which is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Among them, Q, K and V are query, key, and value matrices, and d_k is the scaling factor. To enhance the expressive ability of the model, the multi-head attention mechanism is introduced:

$$\text{MultiHead} = \sum_{i=1}^{h} \text{Attention}(Q, K, V) \tag{2}$$

The model is pre-trained through a large-scale corpus to learn grammatical rules and semantic implication relationships, and has the ability to generate and deduce. The unstructured processing ability of LLM for mental issues and its sensitivity to emotional semantics can provide technical support for context-aware and intelligent auxiliary decision support systems.

2.3. The Combination of Large Language Models and Medical Semantic Modeling

Facing large-scale disease diagnosis and treatment, medical record information is rather complex and unstable, and it is impossible to establish an effective model for the relevant content using traditional processing methods. LLM can be used to mine the meaning of background data and achieve the processing of many medical semantic modeling, such as named body recognition, event extraction, relation extraction, and clinical reasoning [3]. After undergoing customized training for the medical field, LLM is capable of obtaining a series of important elements involved, such as emotional states, behavioral characteristics, and disease vocabulary, and can establish semantic associations related to these elements. And enhance the model's understanding of professional language and causal reasoning ability with the aid of medical ontologies (such as SNOMEDCT, ICD-11) and knowledge graphs. The application of Prompt technology can be used to enable the model to generate contents such as clinical communication, personalized treatment plans and risk prediction [4].

3. Architecture Design and Integration Method of Clinical Decision Support System for Mental Health

3.1. Overall System Architecture Design and Functional Module Division

To construct a modular clinical decision-making mental and psychological support system with good credibility, understandability and suitability for integration in the medical field, the overall framework design of this system includes four layers: Among them, the first layer is the data layer, which is used to receive and process clinical information including structured and unstructured; The second layer is the model layer, mainly including the fine-tuned large language model and the semantic analysis and diagnostic prediction of the combined medical oncology and medical knowledge graph. The third layer is the decision engine layer, which is responsible for Prompt generation, ethical evaluation and result control. The fourth layer is the application layer, which is mainly responsible for visual interaction for physicians and the presentation of personalized suggestions. The operation mode of loose coupling is achieved among each layer through interface services and message relaying, thereby meeting the flexible design requirements of rapid system iteration and module replacement (Figure 1).

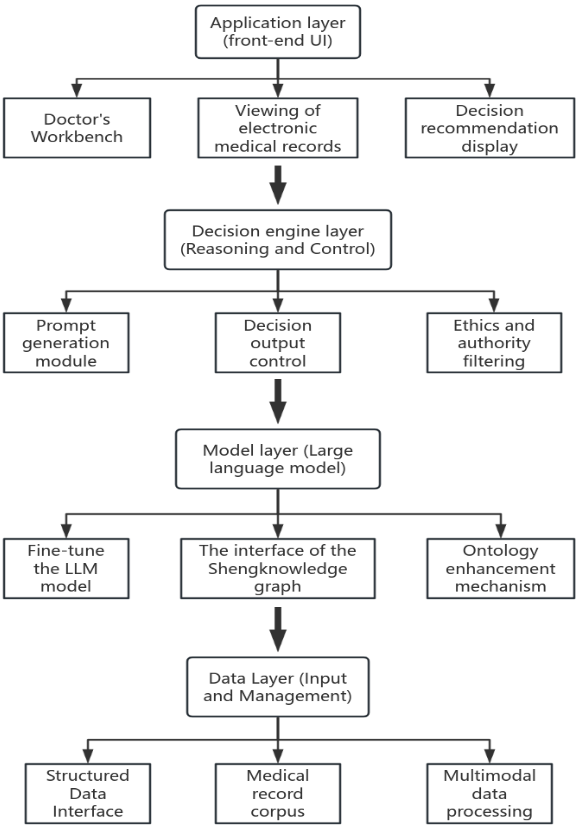
Application layer (front-end UI)
Doctor's Workbench
Viewing of electronic medical records
Decision recommendation display
Decision engine layer (Reasoning and Control)
Prompt generation module
Decision output control
Ethics and authority filtering
Model layer (Large language model)
Fine-tune the LLM model
The interface of the Shengknowledge graph
Ontology enhancement mechanism
Data Layer (Input and Management)
Structured Data Interface
Medical record corpus
Multimodal data processing

Figure 1. Overall architecture of the clinical decision support system for mental health.

3.2. Multi-Source Data Input Processing and Context Modeling Mechanism

In the clinical context of mental health, there are many types and formats of data, including structured (such as specific patient information and test results, etc.), semi-structured (such as specific fields in electronic medical records), unstructured text (such as conversations between doctors and patients or psychological assessment reports), as well as various types of information such as voice or biological signs. In order to support the end-to-end semantic parsing and analysis capabilities, on this basis, a unified data expression specification needs to be established first, and all data should be uniformly operated, processed and transformed based on this specification [5]. The unstructured text after desensitization, word segmentation and syntactic analysis is integrated into the large language model to achieve context encoding. The processed structured data is integrated through sparse vectors or numerical standardization and adjusted to align with the language encoding in the encoding stage.

In terms of modeling strategy, the system adopts a nested context modeling framework based on the multimodal Transformer structure, integrating the input at the current point in time, historical states and external knowledge. Its core formula is as follows:

$$C_t = Enc(X_t, H_{t-1}, K) \tag{3}$$

Among them, X_t represents the current input, which may be text, score, symptom list or other feature vectors; H_{t-1} is the implicit state representation of the previous round, capturing the long-term context; K is the medical knowledge graph or ontology encoding, assisting in the recognition of conceptual relationships. The encoding function

$Enc(\cdot)$ can adopt the BerT-like Encoder or Decoder-only architecture, such as the LLM sub-module. When dealing with multimodal sequences, it uses the shared attention mechanism, allowing different source data to interact at the Token level.

Furthermore, a time encoding module was introduced in this system to discretize the timestamps extracted from clinical records and inject them as an additional one-dimensional coordinate vector into the embedding structure, thereby extracting the time and dynamic information of the data. By using guiding Prompt inputs such as "Could a patient's continuous insomnia for two weeks possibly trigger their risk of depression?" Similar problems can induce the model to complete context analysis and inference with clinical significance.

During the model training stage, for organized annotations, classification penalties (such as interaction entropy) are used for control, while for generated text, language schema penalties are employed. During the testing phase, the system uses a sliding window to maintain the history of multiple conversations. Before that, it undergoes semantic consistency checks in the control section to ensure historical information and correct understanding. The comprehensive strategy takes into account both the recognition ability of phrase modeling for short-term behavioral characteristics and the recognition ability of long-context modeling for long-term disease trajectories, thereby supporting the CDSS application of large language models and achieving stable and accurate understanding of mental health contexts.

3.3. Prompt Engineering and Decision Output Control Strategy

Prompt engineering is a key link in introducing large language models into the clinical decision support system for mental health. Its main purpose is to create a system that can generate meaningful, structured and ethical output content. In the context of mental health, Prompt is not only used to perform natural language tasks, but also can be used as a control entry point for the model, thereby controlling and adjusting its behavior in real time. This model is a hierarchical Prompt construction scheme implemented based on template filling and context connection, with the aid of context understanding, task description and command control. The construction is modular and combinable.

In the implementation of Prompt templates, multiple slots are commonly used, such as: [Patient Summary] + [Task Instruction] + [Output Constraint]. Among them, [Patient Summary] can be automatically generated from medical record information, structured questionnaires, and historical interaction information; [Task Instruction] Extract from the given command set according to the task requirements; [Output Constraint] defines the constraints of the model's output form and lexical space, such as "Please output the JSON structure using DSM-5 terms." The final Prompt forms a continuous text string as the input of the model and uses parameters such as token-level, length, temperature/top-p to control the stability of the model output.

With the aim of increasing the consistency of the output and protecting its security, the Prompt scheduler module is constructed to achieve flexible configuration of template and parameter selection. When this module receives the Task description and status information from the user, it determines the current working stage based on the treatment plan Graph (Task Graph) and the state transition mechanism, and then calls the Prompt route corresponding to this stage. For example, when conducting suicide risk assessment, the Prompt will add instructions for capturing emotional information and include instructions that limit the generation of risk level types, such as: "Please mark your current risk level as one of 'low risk', 'Medium risk', or 'high risk'." Meanwhile, based on the sensitivity of the work, we will reduce the temperature value and the maximum generated token value obtained from the model in real time to avoid illusions or semantic drift in the results.

To meet the usage goals of output management, the system designs a Post-Processing module chain (Post-processing Pipeline) composed of three modules: entity detection, logical consistency, and ethical rule filtering. Among them, the entity detection part checks whether the output results of the model conform to the compliant vocabulary stipulated in the medical database. The consistency judgment re-evaluates the possible outputs by using the probability distribution of the language model itself and excludes low-credibility responses. The ethical filtering stage integrates blacklist words, sentiment orientation analysis, and dialogue scene recognition algorithms to block possible sensitive or misleading information and mark or automatically rewrite it.

3.4. Technical Routes for System Integration and Deployment

To balance inference efficiency, data privacy and the integration degree of personal systems, the system design adopts a distributed microservice architecture, enabling the service interfaces, business logic and data interfaces of the model to operate independently. The deployment includes front-end access, interface portal, central service component and model inference component, and is directly connected with the hospital's information system and service diagram. For the large language model part, the response efficiency can be improved through load balancing and caching with the help of the optimized API service. Different model versions can be arranged as needed. To meet the different compatibility requirements of this part, this system not only provides service approaches such as local installation (for the psychiatric hospital system adapted to this part) and private cloud installation, enabling the model capabilities to be deployed in a closed manner nearby, but also has security guarantees for tracking and recording user operations and data transmission through modules such as log auditing, permission control, and interface password addition. The technical route takes into account scalability and medical-grade stability as a whole, and can be deployed at multiple locations and integrated across systems (Figure 2).

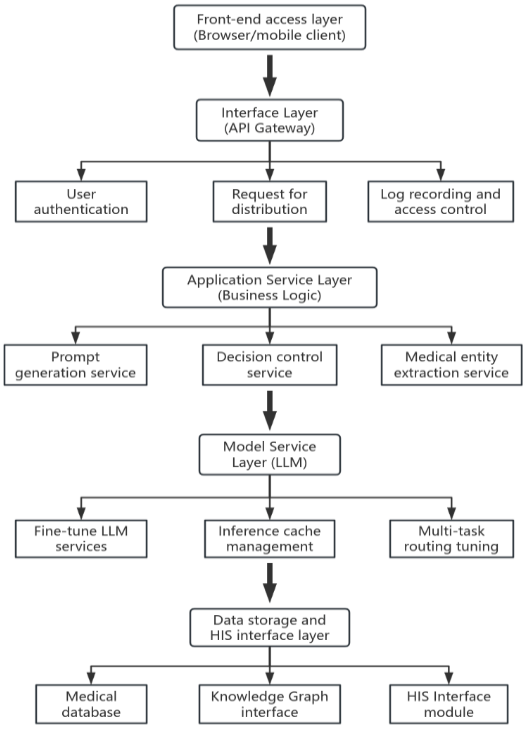
Front-end access layer (Browser/mobile client)
Interface Layer (API Gateway)
User authentication
Request for distribution
Log recording and access control
Application Service Layer (Business Logic)
Prompt generation service
Decision control service
Medical entity extraction service
Model Service Layer (LLM)
Fine-tune LLM services
Inference cache management
Multi-task routing tuning
Data storage and HIS interface layer
Medical database
Knowledge Graph interface
HIS Interface module

Figure 2. Technical framework diagram of system deployment and service integration.

4. Model Evaluation System and Risk Control Mechanism in Mental Health Tasks

4.1. Model Evaluation Index System and Benchmark Dataset Construction

In the clinical decision-making scenarios of mental health, the evaluation criteria for large language models should combine the performance of general NLP and the characteristics of medical tasks to ensure that the models not only generate language naturally but also have sufficient clinical applicability and stability. The overall evaluation score is designed using weighted aggregation indicators and is defined as follows:

$$S = w_1 \cdot \mathrm{BLEU} + w_2 \cdot \mathrm{Acc}_{\mathrm{clinical}} \tag{4}$$

Among them, BLEU measures the degree of language matching, where $\mathrm{Acc}_{\mathrm{clinical}}$ represents the consistency between the model's suggestions and the expert's standard answers, and the two weights w_1, w_2 can be adjusted according to specific tasks.

To support the systematization of evaluation, a multi-task-based large language model mental health benchmark dataset was constructed, and a multi-dimensional evaluation index system was designed based on the task goals.

It can be seen from Table 1 that the tasks of "symptom extraction" and "diagnostic recommendation generation" are where the model most prominently demonstrates its ability to understand and reason about medical semantics. Therefore, the structural accuracy score is selected as the main basis for the evaluation index. The "Emotional state classification" task focuses more on the model's recognition ability in emotional speech features, which can promote the effect of psychological assessment. "Multi-round question and answer coherence" is a technical indicator for measuring how a model remembers and stores information for a long time, which is extremely crucial for CDSS that have the need to handle multi-round dialogues. This assessment system encompasses all performance dimensions ranging from sentence quality to medical efficiency, to single-round and overall understanding, and can conduct a comprehensive evaluation and comparison of the performance of different models in mental health tasks.

Table 1. Evaluation Dimensions and Corresponding Indicators of Mental Health Tasks.

Task type	Sample source	Main evaluation indicators
Symptom extraction	Chief complaint text and dialogue clips	Precision / Recall
Generation of diagnostic suggestions	Medical record summary and past records	Acc_clinical
Classification of emotional states	Psychological interview Record	Emotion-F1
Consistency of multiple rounds of question-and-answer	QA context history	Context-Coherence Score

4.2. Model Robustness and Deviation Detection Mechanism

Due to the uncertainty, polysemity and strong emotionality of clinical texts, in mental health tasks, the robustness of the model under different input perturbations and the identifiability of group discrimination should be focused on and evaluated by using the input perturbation and contrastive response consistency index (CRA):

$$\mathrm{CRA}(x, x') = 1 - \frac{1}{n}\sum_{i=1}^{n} sim(y_i, y_i') \tag{5}$$

Among them, x is the original input, x' is the perturbated sample (semantic substitution, spelling perturbation), y_i, y_i' is the corresponding output, and $sim(\cdot)$ is the semantic similarity function. The lower the CRA value is, the stronger the stability of the model output is.

To detect deviations, we adopt the Group Sensitivity Score (GSS):

$$GSS = \max_{g_1, g_2 \in G} |E[f(x|g_1)] - E[f(x|g_2)]| \tag{6}$$

Among them, G is the set of user groups (such as different genders, ages, and races), and $f(x|g_2)$ is the average response output of the model to the input of a certain group. The larger the GSS is, the more obvious the model deviation is, and further constraints or retraining are required.

4.3. Ethical Risk Control and Explainability Design

In the application of large language models in mental and psychological medical scenarios, the output content contains valuable and sensitive information. The lack of corresponding management measures may lead to incorrect guidance and treatment, biased content, and unethical behaviors. To ensure that the system is reasonable and legal when providing medical guidance, responding to emotions and answering questions, multiple layers of risk management and transparency components should be added to the model's thinking. General risk management can be generally divided into four steps: sensitive word mining, content review, transparent model, and feedback after interaction with people. The steps are divided into three parts, namely model preprocessing, intermediate process control and final analysis and evaluation.

The sensitive content recognition module based on medical ethics adopts the keyword graph as the basis. Through the language model classifier, potential dangerous words, overly strong suggestions or inappropriate semantic expressions in the generated text are identified. If extremely high-risk tags are detected, manual processing paths are adopted or generation is refused. Meanwhile, we use interpretability analysis techniques to compare and calculate the contribution of each input fragment to the final conclusion by using Shapley values and prompts, and present a token-level heat map. This method can track the processing path of the model and the trigger positions of key semantics, and quantify the important semantic locations, enhancing the trust of medical-related practitioners in the system. Meanwhile, in the implementation, support semantic fairness judgment (such as response deviation of multiple populations), evaluate the performance of the model under factors such as gender population and age, and ensure the consistency and ethical unbias of the generation (Figure 3).

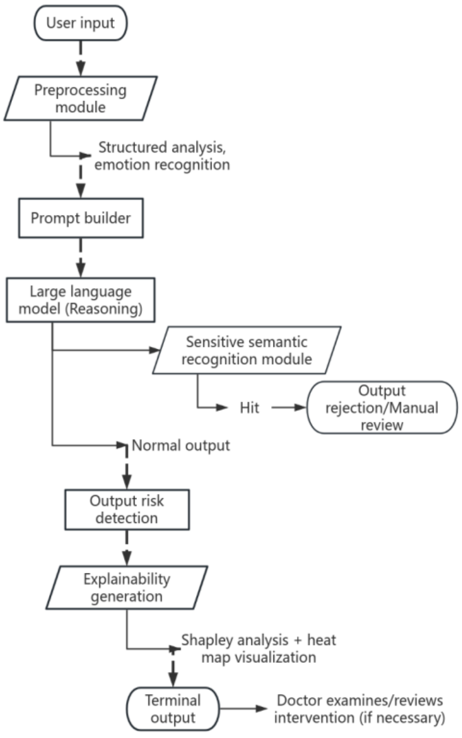
User input
Preprocessing module
Structured analysis, emotion recognition
Prompt builder
Large language model (Reasoning)
Sensitive semantic recognition module
Hit
Output rejection/Manual review
Normal output
Output risk detection
Explainability generation
Shapley analysis + heat map visualization
Terminal output
Doctor examines/reviews intervention (if necessary)

Figure 3. Process of ethics and interpretability control mechanism.

5. Conclusion

The mental health clinical decision support system based on large language models is a new technical route that integrates semantic modeling, Prompt control, multi-source datasets, and interpretable mechanisms. It has great potential in generating diagnostic suggestions, emotion recognition, and risk early warning. Future research focuses will be on enhancing accuracy, adapting to deployment, and improving ethical compliance. Ensure its safety and sustainability in the field of mental health.

References

1. X. Wang, H. Ye, S. Zhang, M. Yang, and X. Wang, "Evaluation of the performance of three large Language models in clinical decision support: A comparative study based on actual cases," *Journal of Medical Systems*, vol. 49, no. 1, p. 23, 2025. doi: 10.1007/s10916-025-02152-9
2. R. Benbenishty, and R. Treistman, "The development and evaluation of a hybrid decision support system for clinical decision making: The case of discharge from the military," *Social Work Research*, vol. 22, no. 4, pp. 195-204, 1998. doi: 10.1093/swr/22.4.195
3. Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large language models for mental health applications: systematic review," *JMIR mental health*, vol. 11, no. 1, p. e57400, 2024. doi: 10.2196/57400
4. X. Yang, T. Li, Q. Su, Y. Liu, C. Kang, Y. Lyu, and Y. Pan, "Application of large language models in disease diagnosis and treatment," *Chinese Medical Journal*, vol. 138, no. 02, pp. 130-142, 2025.
5. Y. Gu, A. E. Andargoli, J. L. Mackelprang, and D. Meyer, "Design and implementation of clinical decision support systems in mental health helpline Services: A systematic review," *International journal of medical informatics*, vol. 186, p. 105416, 2024. doi: 10.1016/j.ijmedinf.2024.105416

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual Aauthor(s) and contributor(s) and not of GBP and/or the editor(s). GBP and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.