Academic Journal of Management and Social Sciences
ISSN: 2958-4396 | Vol. 13, No. 2, 2025

Research on the Construction of Medical Critical Thinking Assessment Gauge Driven by Generative AI

Liang Ying [1, #], Zixun Dai [1, #], Xiaoqing Qiu [2], Zhe Ouyang [2], Yuzhu Pan [1], Xin Fang [1], Jiahe Li [1], Yutong Sun [1], Xiaona Guan [1, *]

[1] Renji college, Wenzhou medical university, Wenzhou, Zhejiang, China
[2] Wenzhou Medical University, Wenzhou, Zhejiang, China

*** Corresponding author:** Xiaona Guan

#These authors are co-first authors of the article

Abstract: With the rapid development of artificial intelligence (AI) technology, major changes have taken place in the field of medical education in China. In recent years, in order to respond to the training requirements of "new medicine" for compound talents, the demand for systematic evaluation of critical thinking ability of medical students in China is increasing. Based on SOAP clinical reasoning framework and integrating existing critical thinking theory, this study established a medical critical thinking assessment gauge covering six dimensions of "interpretation-analysis-evaluation-inference-self-adjustment-clinical adaptation", each dimension has five levels, presenting a path from information processing to clinical decision-making ability, and introducing evidence-based medicine tools (such as AGREE II), cognitive bias and other professional concepts enhance the professionalism and consistency of evaluation, which can be used as the core quantitative basis of the generative AI-driven critical thinking education system. Meanwhile, the gauge realizes the paradigm transformation from static evaluation to dynamic diagnosis and from general scoring to personalized intervention, providing a reliable path for the cultivation of medical high-order thinking ability.

Keywords: Generative AI; Critical Thinking; Measurement Gauge; SOAP Framework.

1. Introduction

Under the background of the construction of "new medicine" in China, the demand for critical thinking of medical students in the field of medical education is increasing day by day. However, although the traditional medical critical thinking assessment tools have certain effectiveness, there are still many limitations, such as specific dynamic scenario simulation based on medical background[1] and other key dimensions showing obvious lag; simulation training mostly stays at the level of single skill operation, lacking system construction for real medical scenarios such as multi-thread diagnosis and treatment tasks and emergency handling[2]; At the same time, the inherent subjectivity of the raters and the closed passive response form of objective questions in the scoring process further weaken the reliability of the evaluation results[3].In order to solve the dilemma of traditional medical education of "emphasizing knowledge and light thinking", artificial intelligence technology has become an important force to solve its long-term pain points[4].Based on this, this study developed a systematic and quantifiable medical critical thinking assessment gauge based on SOAP clinical reasoning framework and international mainstream critical thinking theory, aiming to provide a standardized assessment tool for dynamic assessment supported by generative AI.

2. Core Theoretical Basis

SOAP clinical reasoning framework is a standardized thinking tool verified by long-term clinical practice and education in the medical field. Its name comes from the English abbreviation of the four elements of subjective data, objective data, assessment and plan[5].The framework was initially popularized in medical institutions all over the world as a medical record writing standard, and gradually developed into a structured thinking model supporting clinical decision-making. The logic closed-loop of information collection, comprehensive analysis and scheme formulation ensured the standardization of medical behavior, and disassembled the complex clinical diagnosis and treatment process into four operable and traceable links. Based on the above theoretical support and operational basis, this framework lays a solid foundation for the establishment of a three-dimensional ability system of medical history integration, differential diagnosis and evidence-based decision-making in medical critical thinking assessment[6].

Its core value lies in building a bridge between theoretical knowledge and clinical thinking and practical ability for medical students. Relying on this framework, the critical thinking assessment dimension of medical students can be systematically and quantitatively constructed, forming a standardized paradigm for assessing medical students 'clinical reasoning ability, and providing a standardized thinking assessment template for subsequent integration into the generative AI assessment system[7].

In addition, SOAP framework transforms educational evaluation theory from result-oriented to process-oriented, providing theoretical basis for generative AI to carry out dynamic evaluation. Generative AI transforms the structured thinking process of SOAP framework into quantifiable evaluation dimension, tracks the process data of students diagnosis reasoning and evidence quotation in real time, realizes dynamic evaluation and instant feedback, accurately identifies thinking shortcomings, provides scientific basis for personalized teaching intervention, and effectively breaks

through the limitations of traditional static evaluation.

3. The Establishment of Measurement Model

This study draws on the experience of Richard Paul [8], Diane F. Halpern [9] and others' theoretical definition of the general thinking process has formed a six-dimensional assessment gauge of "interpretation-analysis-evaluation-inference-self-adjustment-clinical adaptation", each dimension can be divided into 5 grades (see Table 1).

Table 1. Critical Thinking Capacity Scale

Dimension and rank	Interpretation	Analysis	Evaluation
5 (Excellent)	It can accurately define the core clinical problems behind complaints, systematically construct chronological medical history including social, family and medication history, and professionally explain the clinical significance of positive and negative signs with differential diagnostic value.	Based on pathophysiological mechanism, it can construct logical causal chain from symptom to diagnosis, actively identify and prioritize critical emergencies that cannot be missed in differential diagnosis, and form a clear diagnosis hypothesis system.	Ability to critically assess guidelines and literature using evidence-based medicine tools (e.g. AGPEI 1), pinpointing limitations in study design, population applicability, and demonstrating how well the evidence fits current patient context and decision weight.
4 (Good)	Accurate identification of major symptoms and key elements of medical history, standardized description of important positive/negative signs, but may miss some social psychological factors or atypical medical history collection.	It can establish reasonable relationship between symptoms and underlying mechanisms, and the differential diagnosis framework covers the main differential directions. The logic is basically self-consistent, but the breadth of disease spectrum or complex etiology are not fully considered.	Ability to identify major types of bias in studies (e.g. selection bias), basic judgment on the level of evidence, but depth and flexibility in applying evidence individually to clinical decisions.
3 (Qualified)	The understanding of complaints is biased or information missing, the framework of medical history collection is incomplete, the description of physical signs stays at the level of identification and listing, and lacks deep correlation with disease physiology.	Symptom analysis mostly presents isolated possibility list, shallow explanation of mechanism, narrow differential diagnosis range or incoherent logic chain, lack of systematicness.	Only "for/against" binary judgment of evidence, insufficient understanding of research methodology, lack of systematic critical thinking framework for evaluation process.
2 (Poor)	Ambiguous grasp of complaints, scattered and non-sequential medical history information, important omissions or misreadings of key signs (such as specific signs and vital signs).	Symptom analysis lacks systematic path, pathophysiological basis is weak, diagnosis basis provided lacks effective logical association, inference process has obvious jump.	Evidence judgment relies heavily on personal experience or intuition, fails to recognize obvious flaws in the study design, and the evaluation process is subjective and arbitrary.
1 (Not Qualified)	Inability to accurately understand patient complaints, serious missing or disjointed history collection, principled errors or confusion in description of basic signs.	Failure to establish meaningful links between symptoms and pathophysiology, lack of basic clinical analysis skills, and confusion in diagnostic thinking.	They do not have basic ability to judge evidence, cannot distinguish between research results of different quality, and are completely inaccurate in evaluation.
Dimension and rank	Inference	Self-Regulation	Clinical Adaptation
5 (Excellent)	Based on incomplete, conflicting, or dynamically evolving clinical information, a differential diagnosis checklist sorted by urgency and likelihood can be generated, and individualized, full-process management protocols including diagnostic treatment, monitoring indicators, and follow-up plans can be developed.	It can continuously monitor metacognition during the whole process of clinical reasoning, actively identify and strategically correct cognitive biases (such as anchoring effect and verification bias), and demonstrate efficient thinking optimization ability based on new evidence and reflection.	The diagnosis and treatment scheme is an optimal solution that achieves high individualization within the framework of authoritative guidelines, perfectly balances efficacy, safety, cost-effectiveness and patient value orientation, and has clear feasibility demonstration for realistic medical resources and policy environment.
4 (Good)	Diagnostic thinking consistent with standard clinical pathways can be developed, treatment options are safe and have core evidence-based support, but consideration of comorbidity management, treatment options, or long-term prognosis is insufficient.	Can detect obvious contradictions or uncertainties in reasoning and try to adjust them, but the initiative, systematicness or efficiency of correction needs to be improved.	The protocol conforms to the diagnosis and treatment norms and ethics, considering the main clinical indicators and basic conditions of patients, but needs to be further improved in health economic evaluation or precise individualized adaptation.
3 (Qualified)	The proposed diagnostic assumptions tend to be conventional and lack of hierarchy, the treatment plan is "standardized" and lacks flexibility, and the response plan for treatment failure or sudden change is insufficient.	Occasionally aware of knowledge blind areas or thinking limitations, but lack of systematic improvement methods and in-depth reflection, correction behavior is relatively passive.	The protocol followed basic medical principles but had significant deficiencies in individualization, inclusion of patient preferences, or multidisciplinary collaboration.
2 (Poor)	There is a serious disconnect between diagnostic assumptions and clinical manifestations, treatment options lack critical evidence-based support or logical contradictions, and there are fundamental gaps in clinical reasoning.	Seldom examine their own diagnostic thinking actively, even if they find mistakes under external prompts, they lack effective and feasible adjustment strategies, and their thinking tends to solidify.	The protocol is seriously disconnected from the clinical actual situation, and the operational safety, potential complications and ethical risks are not considered enough, and the feasibility is low.
1 (Not Qualified)	It is difficult to form a reasonable clinical reasoning chain, the proposed diagnosis and treatment scheme is seriously inconsistent with the patient's condition, and even there are principled errors or ethical risks.	Complete lack of self-monitoring awareness, rigid thinking, inability to identify and correct obvious logical errors or factual deviations.	The protocol violates basic medical principles and ethical norms, has serious potential safety hazards, and has no clinical implementation value.

Among them, the dimension of "interpretation" focuses on students' accurate understanding and clear expression ability of complaints, medical history and objective signs; The dimension of "analysis" focuses on whether students can effectively establish the logical association between symptoms and pathophysiological mechanisms; The dimension of "evaluation" aims to examine students' ability to critically discriminate the quality of evidence and the reliability of sources; The dimension of "inference" is used to measure students' ability to generate reasonable diagnosis and make treatment plan under limited information conditions; The dimension of "self-regulation" reflects students' ability to identify errors and correct their thinking in the process of reasoning.Since the ultimate goal of medical education is to train qualified clinical decision makers, this study adds a new dimension of "clinical fitness" to the original framework. This dimension focuses on evaluating whether the diagnosis and treatment plan formulated by students has practical feasibility, safety and individualization degree. Specific consideration factors include: whether the latest guidelines and evidence-based principles are strictly followed, whether the patient's age, comorbidities and social psychological background are fully considered, whether the balance judgment between risk and benefit is made, and whether the implementation feasibility of the plan under the actual medical resource conditions[10]. The introduction of this dimension makes the model closer to the new medical education orientation.

Each dimension of this scale is deeply integrated with SOAP clinical reasoning structure, and its hierarchical description implies an advanced thinking path from information processing relationship construction to critical discrimination and decision optimization. For example, the "analysis" dimension clearly defines the ability ladder from "listing possibilities"(level 3) to "establishing logical connections"(level 4) to "constructing complete causal chains and identifying acute and severe diseases"(level 5), thus enhancing the discrimination and operability of the scale.At the same time, the scale clearly introduces specific concepts such as evidence-based medical tools (such as AGREE II), differential diagnosis of critical emergencies and cognitive bias (such as anchoring effect), so that students and raters (including AI system) can judge according to professional and clear standards, thus significantly improving the professionalism and consistency of evaluation. Especially prominent is that the dimension of "clinical adaptation", as the innovation core and foothold of this scale, not only pays attention to the correctness of diagnosis and treatment scheme, but also devotes itself to seeking the optimal balance among authoritative guidelines, individual needs of patients, cost-benefit, values and realistic medical resources, reflecting the advanced concept of value-oriented medical treatment. At the presentation level, the scale adopts strict academic terms and concrete action verbs (such as "build","explain", "identify" and "discuss"), effectively avoiding colloquialism and vague expression, so as to make the description structure clear and the direction clear.This kind of design is not only conducive to the accurate understanding of scoring standards by generative artificial intelligence, but also provides accurate reference for teachers to carry out manual evaluation, thus strongly supporting the project idea of "man-machine collaborative double scoring mechanism."

4. The Integration Path of Gauge in Generative AI Evaluation System

This gauge is the core evaluation basis of generative AI medical critical thinking evaluation system. Its integration path includes semantic alignment of structured gauge and AI reasoning and closed-loop evaluation from data input to multidimensional feedback.

(1) Semantic alignment of structured gauge and AI reasoning

In order to achieve effective evaluation of critical thinking process by generative AI, the six dimensions of interpretation, analysis, evaluation, inference, self-regulation and clinical adaptation covered by this gauge, as well as the five-level behavior description under each dimension, all adopt the "cue-score" mapping system that AI can recognize. When AI parses the text input by students, such as medical history collection, diagnostic hypothesis and diagnosis plan, according to the keywords set in the gauge, Logic structure and behavior standard, carry out multi-round semantic matching and reasoning chain analysis, and then realize quantitative scoring of thinking process according to dimensions.

(2) losed-loop evaluation from data input to multidimensional feedback

First, the intelligent embedding of gauges.Through fine-tuned medical large model and natural language processing technology, the system converts the six dimensions of the gauge into computable analysis units.In interactive case analysis, the system will parse students text input in real time, automatically identify key behavior nodes corresponding to each dimension of the gauge, such as "identifying critical emergencies" and "evaluating evidence levels", and realize fine tracking of thinking process.

Second, dynamic scoring and path visualization.Based on the identified behavior nodes, the system automatically scores 1-5 points for each dimension according to preset gauge rating criteria.At the same time, the graph structure contrast algorithm is used to visually compare the real-time reasoning path of students with the standard path of experts.This visualization not only visually presents the trajectory of thinking, but also clearly indicates logical deviations and cognitive patterns that are difficult to detect with traditional methods[11].

Third, closed-loop feedback and personalized intervention. The system automatically generates structured diagnostic feedback reports by locating thought biases.More importantly, it can accurately push relevant learning resources and training cases according to the ability shortcomings revealed in the report (i.e. the advantages and disadvantages of students), thus forming a complete "evaluation-feedback-improvement" teaching closed loop, realizing accurate cultivation and effective promotion of critical thinking.

5. Conclusion

Based on SOAP clinical reasoning framework, a six-dimensional and five-level evaluation scale was developed, which transformed abstract clinical thinking process into quantifiable indicators and provided a standardized tool for evaluating critical thinking of medical students. The scale innovatively introduced the dimension of "clinical fit", emphasizing that diagnosis and treatment should balance

evidence-based principles, individual needs of patients and practical medical resources.Patient-centered medical service mode is the core of value-oriented medical service mode, and evaluating medical service quality from the perspective of patients is an innovative practice of modern medical service quality management. This gauge not only provides a standardized tool for training and evaluating critical thinking of medical students, but also provides a structured evaluation template for the application of generative AI in educational evaluation. It also reflects the core concept of value-oriented medicine, which has important theoretical value and practical significance.

Acknowledgments

Funding Project: Annual *Projects of Zhejiang Provincial Philosophy and Social Science Planningfor the Year 2026, Project Number: 25WSK009YB.*

References

[1] Cui Liyuan, Zhu Yaxin, Qu Bo.Application and prospect of critical thinking scale in medical education [J].Fudan Education Forum, 2021, 19 (02):106-112.DOI:10. 13397/j. cnki. fef.2021.02.016.

[2] Wartman S A. Medicine, machines, and medical education[J]. Academic Medicine, 2021, 96(7): 947-950.DOI:10. 1097/ ACM. 0000000000004113.

[3] Jiang Liming, Liu Yujie, Luo Fang.Critical Thinking Assessment Based on Real Problem Situations: Status and Challenges [J].China Distance Education, 2022,(12):58-67+ 77+ 83. DOI:10.13541/j.cnki.chinade.2022.12.006.

[4] Zhang Kehui, Zhang Wei.Evolution and Trend Analysis of Medical Artificial Intelligence Research Hotspots at Home and Abroad [J/OL]. Hygiene Soft Science, 1 -8[2025-10-13]. https:// link.cnki.net/urlid/53.1083.R.20250916.1721.017.

[5] Huang J, Zhang H, Fan X S, et al. Application of SOAP teaching method combined with Mini-CEX in outpatient medical record writing teaching of standardized training for dental residents [J].Zhejiang Medical Education, 2023, 22 (04): 237-241. DOI: 10.20019/j.cnki. 1672-0024.2023. 04. 237. 05.

[6] Naqvi W M, Ganjoo R, Rowe M, et al. Critical thinking in the age of generative AI: implications for health sciences education [J]. Frontiers in Artificial Intelligence, 2025, 8: 1571527.

[7] Calm down, Lu Honghuan, Dailin.Generative AI Enabling Critical Thinking Assessment--An Applied Experiment Based on ChatGPT [J]. Modern Distance Education Research, 2024, 36 (06): 102-111.DOI:CNKI: SUN:XDYC.0.2024-06-011.

[8] Richard Paul, Linda Elder.Critical Thinking Tools (3rd Edition) [J].Enterprise Observer, 2018,(09):120-121. DOI:CNKI: SUN: QYGC. 0.2018-09-051.

[9] F. D H. Thought and Knowledge: An Introduction to Critical Thinking[M]. Taylor and Francis:2013-11-07: DOI: 10.4324/ 9781315885278.

[10] Calm down, Lu Xiaoxu.A Study on Critical Thinking Ability of Item Bank Game Evaluation [J].Open Education Research, 2020, 26 (01):82-89. DOI:10.13966/j.cnki.kfjyyj.2020.01.009.

[11] Liu Wei, Zhou Ning, Zhang Fangfang.Research on text-based information visualization method [J].Modern Library and Information Technology, 2003,(02):34-36+ 47. DOI:CNKI: SUN: XDTQ.0.2003-02-009.

[12] Li Qiaojun, Yan Jin, Zhang Zhonglin, et al. An Empirical Study on Core Evaluation Index System of Hospitalization Service Quality Based on Value Care [J]. China Health Economics, 2020, 39 (08):79- 80. DOI: CNKI:SUN:WEIJ.0.2020-08-023.