Viewpoint

# From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era

Weiping Cao[1,2], MMed; Qing Zhang[1*], MD; Jialin Liu[3,4*], MD; Siru Liu[5], PhD

[1]Department of Cardiology, West China Hospital, Sichuan University, Chengdu, Sichuan, China

[2]Department of Cardiology, The People's Hospital of Leshan, Leshan, Sichuan, China

[3]Information Center, West China Hospital, Sichuan University, Chengdu, China

[4]Department of Medical Informatics, West China Medical School, Sichuan University, Chengdu, China

[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[*]these authors contributed equally

**Corresponding Author:**
Jialin Liu, MD
Information Center
West China Hospital
Sichuan University
No.37 Guo Xue Xiang Street
Chengdu, 610041
China
Phone: 86 28 85422416
Fax: 86 28 85422607
Email: dljl8@163.com

## Abstract

Large language models are rapidly transitioning from pilot schemes to routine clinical practice. This creates an urgent need for clinicians to develop the necessary skills to strike the right balance between seizing opportunities and taking accountability. We propose a 3-tier competency framework to support clinicians' evolution from cautious users to responsible stewards of artificial intelligence (AI). Tier 1 (foundational skills) defines the minimum competencies for safe use, including prompt engineering, human–AI agent interaction, security and privacy awareness, and the clinician-patient interface (transparency and consent). Tier 2 (intermediate skills) emphasizes evaluative expertise, including bias detection and mitigation, interpretation of explainability outputs, and the effective clinical integration of AI-generated workflows. Tier 3 (advanced skills) establishes leadership capabilities, mandating competencies in ethical governance (delineating accountability and liability boundaries), regulatory strategy, and model life cycle management—specifically, the ability to govern algorithmic adaptation and change protocols. Integrating this framework into continuing medical education programs and role-specific job descriptions could enhance clinicians' ability to use AI safely and responsibly. This could standardize deployment and support safer clinical practice, with the potential to improve patient outcomes.

***(J Med Internet Res 2026;28:e86550)*** doi: 10.2196/86550

**KEYWORDS**

large language model; clinician; agent; competency; artificial intelligence; education; continuing medical education

## Introduction

The convergence of generative artificial intelligence (AI) and health care is catalyzing a paradigm shift in clinical practice, with significant implications for the future of medicine [1-3]. Large language models (LLMs), exemplified by recent advances, such as GPT-4 and Gemini, demonstrate a transformative capacity to process multimodal data and generate context-aware responses, increasingly positioning them as integral components in frontline clinical decision support [4,5].

Although LLMs have the potential to improve clinical effectiveness, ensuring that their application optimizes patient safety, ethical alignment, and long-term benefits remains a substantial challenge [2,5]. This complexity is compounded by the intersection of regulatory requirements and ethical obligations. Evolving legal frameworks, such as the European Union (EU) AI Act and the US Food and Drug Administration

(FDA) guidance, explicitly mandate human oversight for high-risk AI systems [6,7]. Simultaneously, global ethical standards from the World Health Organization (WHO) and the American Medical Association emphasize the necessity of physician leadership and accountability [2,8,9]. However, a gap remains in translating these high-level mandates into actionable clinical skills. Without the active leadership and input of clinicians, these technologies risk imposing unintended burdens and may fail to achieve their full potential [1].

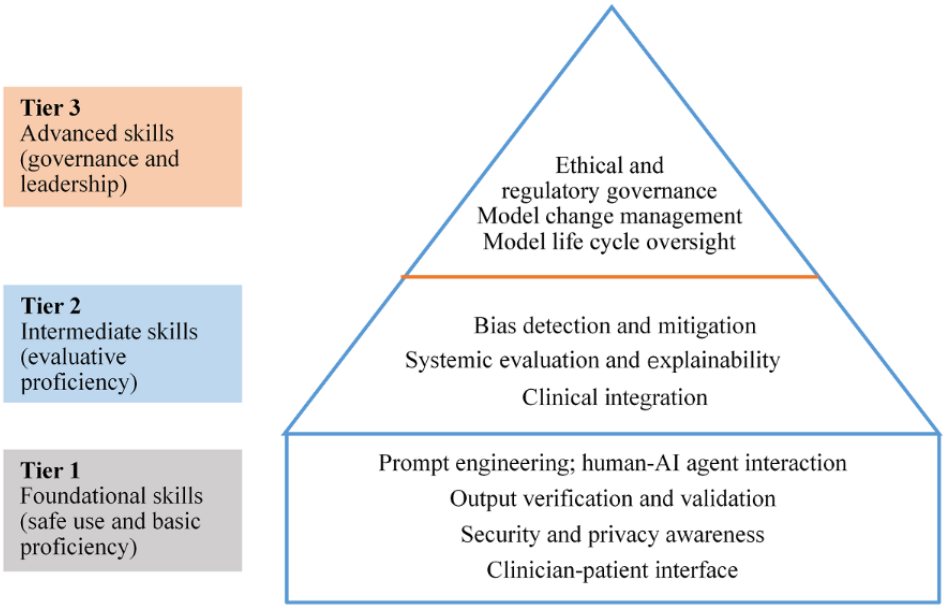
The imperative for advanced AI governance arises from a fundamental shift from passive information retrieval to autonomous task execution. While conventional LLM paradigms rely on user-initiated prompt response exchanges, clinicians query the model and verify its text outputs, and the system does not autonomously call external tools. By contrast, emerging agentic workflows introduce a perceive-plan-act (and often reflect) loop [10]. In this mode, the system interprets high-level clinical intents (eg, hypertension management); decomposes them into subtasks; and autonomously executes actions via application programming interfaces, such as accessing electronic health record (EHR) data or calculating risk scores [11,12]. This transition reframes supervision; clinicians must move beyond prompt engineering to govern how autonomy is delegated, how actions are constrained, and how escalation pathways are formalized.

To address these regulatory, ethical, and technical demands, we propose a foundational, tiered AI competency framework for clinicians. The framework is structured around progressive tiers: tier 1 (foundational skills), tier 2 (intermediate skills), and tier 3 (advanced skills). We describe the core competencies at each tier, outline the framework's limitations, and propose priority directions for validation to sustain its relevance amid an evolving regulatory landscape.

## AI Competency Framework

LLM-enabled care necessitates a transition in the roles of clinicians (physicians, nurses, pharmacists, and allied health professionals)—from interpreting predictive outputs to supervising agentic workflows. Drawing on previous research, a narrative synthesis of evolving digital health competencies, and an analysis of the technical capabilities of LLMs [13-17], we propose a 3-tier, governance-aligned framework that articulates core LLM competencies. As illustrated in Figure 1, the framework progresses from foundational safe use (tier 1) to evaluative proficiency (tier 2) and ultimately to governance and leadership (tier 3).

**Figure 1.** Essential artificial intelligence (AI) skills for clinicians in the large language model era.




### Tier 1: Foundational Skills (Safe Use and Basic Proficiency)

These entry-level competencies prioritize basic interaction with LLMs, enabling clinicians to leverage AI for routine tasks without compromising clinical autonomy. The key elements are described subsequently. First, prompt engineering (task specification for clinician-initiated and hybrid workflows) is used to craft precise, context-aware instructions—with explicit roles, required inputs, constraints, task steps, and output formats (including citation and traceability requirements)—to elicit task-appropriate outputs (eg, structured outlines for differential diagnosis). This competency primarily supports clinician-initiated chat and hybrid workflows, as fully agentic perceive-plan-act execution is typically governed by system-level prompts and policies rather than user-generated prompts. When paired with verification and source grounding, this approach may reduce hallucinations and improve relevance and completeness [18]. Second, human-AI agent interaction (agent supervision) ensures that agents operate within bounded autonomy with explicit roles, goals, and guardrails. Clinicians must maintain awareness of least privilege tool permissions and system constraints (eg, data minimization, time and step limits, and sandboxed execution) [19], with clear termination and escalation criteria. Clinicians also monitor and validate the perceive-plan-act-reflect loop using provenance and citation requirements; protected health information redaction; and audit

logging of prompts, tool calls, and human overrides [20-22]. When confidence or calibration thresholds are not met (eg, coverage targets and abstention or deferral rules), clinicians intervene, interrupt the agent, or revert to manual workflows and document the event for review. Third, output verification and validation involve clinicians critically evaluating individual LLM outputs for accuracy, relevance, and internal consistency. Generated content (eg, summaries, diagnostic considerations, and treatment plans) is cross-referenced against the EHR, structured data, and established clinical evidence to detect hallucinations, omissions, or misstatements. In culturally diverse settings, clinicians must assess outputs for cultural safety and linguistic accuracy. This involves verifying that translated instructions and culturally specific dietary or lifestyle advice are appropriate for the patient's context. Clinicians should also check for Anglocentric bias that could conflict with local norms or the patient's language proficiency. This human-in-the-loop verification is essential for ensuring patient safety in individual clinical encounters [23,24]. The fourth element is security and privacy awareness. To comply with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation, clinicians must adhere to foundational safeguards centered on data minimization and appropriate tool use. For nonintegrated or open access LLM interfaces, this includes avoiding entry of protected health information and direct identifiers and, when clinically necessary to discuss a case, removing or generalizing nonessential identifiers before input [2,5,25,26]. In contrast, for authorized, integrated enterprise agents operating within a secure EHR environment, manual deidentification is often neither feasible nor necessary; instead, clinicians verify minimum necessary access, confirm the agent is scoped to the correct patient context, and ensure permissions are aligned with the specific clinical task through role-based access control and least privilege settings. Rather than conducting technical audits themselves, clinicians prevent inadvertent privacy breaches by distinguishing approved tools from unapproved ones and escalating permission or access-scope concerns through institutional channels. These baseline competencies are prerequisites for safe AI use in routine clinical workflows [5]. Fifth, clinician-patient interface (transparent communication and shared decision-making) involves incorporating AI-assisted content into the clinical encounter without undermining patient trust or the therapeutic alliance. Clinicians should disclose when AI is used (eg, AI-scribed summaries, patient-portal messages, and patient education materials) to uphold patient autonomy and informed consent [26].

### Tier 2: Intermediate Skills (Evaluative Proficiency)

Building on foundational knowledge, these competencies center on critically assessing and integrating LLMs into clinical reasoning workflows while addressing bias and uncertainty in generative AI. First, bias detection and mitigation require clinicians to interpret algorithmic bias audit reports and uncertainty quantification outputs (eg, confidence intervals, prediction intervals, or conformal prediction sets when provided) to assess reliability across patient subgroups. Clinicians initiate and validate remediation actions—such as recommending prompt or workflow adjustments and defining escalation and deferral rules—in coordination with technical teams, ensuring adherence to prespecified fairness metrics and minimum subgroup performance thresholds [27,28]. For example, in tumor grading, clinicians review reported subgroup performance using minimum sample-size thresholds, calibration and coverage, abstention rates, and uncertainty displays (including confidence sets). They assess model rationale and interpret between-group performance gaps. Second, systemic evaluation and explainability involve moving beyond checking individual outputs to assessing the broader reliability, calibration, and failure modes of the AI system. Clinicians should be able to interpret model performance metrics (eg, sensitivity, specificity, error and hallucination rates, and performance in specific subpopulations) and evaluate available explainability outputs (eg, feature importance scores, reason codes, or saliency maps where available) to understand why a model reached a conclusion [29]. This evaluation must include equity audits that assess model performance across distinct subgroups (eg, race, ethnicity, and language) [28]. For instance, a model may demonstrate high overall accuracy but fail disproportionately for languages spoken by minority groups or specific dialects. Clinicians leading the evaluation must identify such disparities and determine if the model is safe for deployment in diverse populations. These skills enable clinicians to judge systemic trustworthiness and identify appropriate clinical use cases and target populations for which the model is calibrated, effective, and equitable [30]. Third, clinical integration requires clinicians to use domain knowledge to refine model outputs (eg, align treatment suggestions with evidence-based guidelines) while monitoring for potential deskilling. Clinicians maintain human-AI collaboration and specify deferral and escalation rules (eg, abstention thresholds and human-review triggers) and document these events for auditability [31,32].

### Tier 3: Advanced Skills (Governance and Leadership)

Unlike the foundational skills in tiers 1 and 2, this tier represents a specialized track for clinician-leaders, clinical informaticists, and physician builders assuming governance roles. These competencies focus on the strategic oversight and architectural direction of AI systems. The main competencies involved are described subsequently. First, ethical and regulatory governance involves overseeing the development of institutional policies for LLM use to ensure alignment with ethical principles, data protection laws (eg, General Data Protection Regulation and HIPAA), and international guidance [26]. This requires establishing governance infrastructure—such as AI steering committees and ethics review boards—to specify authorized use cases, roles and responsibilities, liability frameworks, and compliance protocols. Crucially, policies must explicitly delineate accountability boundaries among supervising clinicians, health care institutions, and AI developers and vendors, particularly for autonomous or semiautonomous agentic workflows. In this capacity, clinician leaders do not personally conduct technical audits; instead, they serve as the strategic link between medical staff and technical bodies, ensuring that institutional processes reflect clinical realities and patient safety risks. Second, model change management requires supervision of domain adaptation (eg, task- or specialty-specific tuning) within a multidisciplinary process. In this capacity, clinicians

bridge clinical needs and technical implementation, upholding standards for validity, equity, and safety. This supervision necessitates predefined evaluation plans, comprehensive documentation (eg, model cards), and rigorous external validation (including multicenter, temporal, and geographic shift tests) before production deployment. Leaders must specify minimum performance thresholds and mandate shadow deployment phases to validate safety before full patient exposure [33,34]. Third, model life cycle oversight entails governing AI systems across their full life cycle—from validation through postmarket monitoring, updating, and decommissioning. This includes orchestrating institutional processes for drift detection, performance re-evaluation, and version control [35,36] (Textbox 1). Leaders must navigate complex regulatory mechanisms for iterative improvement, such as the predetermined change control plans (PCCPs) by the FDA [6] and the postmarket surveillance requirements of the EU AI Act [7]. They collaborate with informatics, regulatory, and quality teams to ensure that updates, retraining, or expanded indications are clinically justified, transparently communicated, and supported by robust evidence and incident review protocols [35,37].

**Textbox 1.** Clinical vignette—governance in action.

**Scenario: executing a manual rollback protocol**

- A clinical informatics director oversees a deployed discharge-summary agent. During routine postmarket surveillance, the monitoring dashboard signals that the model's summarization accuracy has dropped below the prevalidated threshold of 95% (a metric specified in the Food and Drug Administration–accepted predetermined change control plan). Attributing the decline to data drift caused by a recent update in the hospital's note-template format, the director initiates a rollback protocol—leveraging either institutional version control or a vendor-mediated pathway specified in the service-level agreement. The deployment is rolled back to the previous stable version (version 2.0) while the technical team remediates and revalidates the updated model (version 2.1).
- In deployments where direct rollback is technically unsupported (eg, some software as a service–based integrations), the protocol mandates pausing or disabling the agent and reverting to manual workflows until remediation is complete.

## Alignment and Differentiation From Existing Frameworks

This framework is broadly aligned with the American Medical Association's guidance on augmented intelligence, prioritizing physician leadership, transparency, and patient benefit [38,39]. Furthermore, it adheres to competency-based digital education frameworks from the WHO and the Association of American Medical Colleges, both of which prioritize observable behaviors and measurable learning outcomes [40-42]. It also builds on recent competency proposals in AI and digital health that foreground digital health literacy, awareness of data bias, and the ethical use of assistive tools [10,43]. As summarized in Table 1, our contribution lies in extending these earlier frameworks to the agentic LLM era. First, we explicitly differentiate between predictive and informational paradigms and agentic workflows. Accordingly, we move from clinicians interpreting decision support outputs to supervising and governing active, tool-using agents. Second, we introduce model life cycle literacy as an explicit competency domain, encompassing familiarity with mechanisms for ongoing monitoring, updating, and regulatory adaptation. Within this broader, jurisdiction-agnostic concept, PCCPs in the FDA context are presented as one concrete example, alongside emerging requirements under frameworks, such as the EU AI Act. To our knowledge, previous frameworks have not explicitly integrated agent supervision and life cycle–oriented governance into a tiered, clinician-facing competency model.

**Table 1.** Comparison of the agent to the governance framework and existing digital health competency frameworks.

| Feature and dimension | Agent to governance framework | Existing frameworks (eg, World Health Organization, American Medical Association, and Association of American Medical Colleges) |
|---|---|---|
| Technological scope | Agentic and autonomous: agentic workflows (perceive-plan-act loops) and tool-using large language models that execute multistep tasks | Predictive and informational: clinical decision support, diagnostic classifiers, and standard information retrieval systems |
| Clinician's role | Supervisor and governor: human-on-the-loop oversight for task delegation, monitoring agent behavior, and managing bounded autonomy | Interpreter and decision-maker: human-in-the-loop integration, focusing on the critical appraisal of risk scores and diagnostic suggestions |
| Verification skills | Output verification and logic checking: detection of hallucinations in generative text and verification of agentic tool calls (eg, application programming interface actions) | Statistical and evidence-based appraisal: evaluation of model performance metrics (eg, sensitivity and specificity), data quality, and automation bias |
| Regulatory and life cycle | Life cycle management: specific literacy in predetermined change control plans, algorithmic drift detection, and postmarket surveillance (eg, European Union Artificial Intelligence Act and Food and Drug Administration) | Foundational ethics and compliance: adherence to core bioethical principles (beneficence and equity), privacy standards (Health Insurance Portability and Accountability Act and General Data Protection Regulation), and informed consent |
| Target audience and structure | Tiered differentiation: distinguishes between frontline users (tiers 1 and 2) and a specialized leadership track (tier 3) for governance | Universal digital literacy: baseline digital health competencies applicable to the broad health care workforce to ensure safe general use |

## Operationalizing Competencies for Education and Assessment

Translation of this framework into continuing medical education (CME) curricula requires the specification of observable, assessable behaviors aligned with competency-based medical education principles. Given that the clinical workforce encompasses diverse roles—including physicians, nurses, and allied health professionals—implementation and assessment should be tailored to role-specific scope of practice and role-based EHR access controls. For example, behavioral indicators involving least privilege enforcement or the rejection of agent actions may be operationalized differently depending on the individual's credentialed permissions and administrative privileges. To ensure implementation feasibility and mitigate workforce burden, only tier 1 competencies are intended for the general clinical workforce, whereas tiers 2 and 3 are reserved for smaller groups of superusers and clinician leaders in formal governance roles. To avoid adding entirely new courses, these competencies are designed to be integrated into existing curricula (eg, evidence-based medicine, clinical reasoning, and quality and safety) and CME activities. Institutions are responsible for resourcing and coordinating these training activities, ensuring that individual clinicians are not expected to acquire advanced competencies (tiers 2 and 3) without appropriate organizational support and protected time.

Table 2 links each tier to behavioral indicators written as active, measurable learning outcomes. Indicators span tier 1 (eg, identifying hallucinations) to tier 3 (eg, initiating life cycle protocols) and should be tailored to role-specific responsibilities and decision rights. These indicators provide curriculum developers with a concrete scaffolding to design simulation-based, workplace-based, and microlearning assessments that verify skill acquisition in clinical practice. Ultimately, these anchors facilitate the incorporation of this framework into CME curricula and clinical job descriptions, thereby promoting institutional transparency, accountability, and regulatory alignment [40,43,44].

**Table 2.** Sample behavioral indicators for continuing medical education assessment and clinical application.

| Core competency | Behavioral indicator (observable action) |
|---|---|
| **Tier 1: foundational (frontline user)** | |
| Prompt engineering | Formulates a context-aware prompt that includes explicit role definitions (eg, act as a cardiologist), constraints, and required output formats, without disclosing PHI[a] |
| Human-AI[b] agent interaction | Identifies and intercepts inappropriate agent requests (eg, social history for refills) and enforces denial or escalation protocols based on least privilege guardrails and predefined termination and handoff criteria |
| Output verification and validation | Detects and corrects a hallucinated reference or dosage in a large language model–generated draft by cross-referencing with the patient's structured laboratory data and trusted guidelines |
| Security and privacy awareness | Uses minimum necessary data; deidentifies data for nonintegrated tools; for electronic health record agents, verifies patient context and least privilege access, and escalates PHI or policy risks |
| Clinician-patient interface | Informs patients when AI is used, explains AI-derived insights in patient-appropriate language (including uncertainties and limitations), and documents consent or refusal when clinically indicated. |
| **Tier 2: intermediate (superuser or champion)** | |
| Bias detection and mitigation | Interprets stratified subgroup performance and uncertainty reports, flags clinically meaningful disparities, triggers mitigation (eg, threshold adjustments or human-review rules), and verifies improvement via updated audit reports |
| Systemic evaluation | Evaluates a confusion matrix for a diagnostic AI tool to determine if the false-negative rate is acceptable for a specific screening population |
| Explainability | Interprets available explainability outputs (eg, feature importance, reason codes, or saliency maps where available) to detect spurious cues and document potential failure modes |
| Clinical integration | Defines where AI outputs enter the workflow; assigns roles, documentation, and escalation steps; and maintains clinician accountability when AI recommendations conflict |
| **Tier 3: advanced (governance leader)** | |
| Ethical and regulatory governance | Drafts and implements an institutional policy that establishes escalation pathways and explicitly delineates accountability and liability boundaries among the supervising clinician, the institution, and the AI developer for autonomous agentic workflows |
| Model change management | Initiates and justifies model change requests (eg, recalibration, retraining, or expanded indication), defining the clinical rationale, validation plan, and monitoring criteria consistent with the predetermined change control plan |
| Model life cycle oversight | Oversees monitoring of model performance and drift (eg, calibration, error rates, and data shifts); ensures execution of incident protocols and predefined controls (eg, roll back and human review) |

[a]PHI: protected health information.

[b]AI: artificial intelligence.

## Limitations and Future Work

We propose a governance-aligned competency framework designed to guide clinicians in the safe and effective use of LLMs in clinical practice. However, several limitations should be acknowledged. First, external validity may differ by specialty, care setting (inpatient vs ambulatory), health-system maturity, and EHR integration capacity. Critically, the institutional infrastructure required for tier 3—specifically, the establishment of AI steering committees—may currently be feasible only in resource-rich academic medical centers. Mandating such governance structures in resource-constrained community hospitals may be impractical. This feasibility gap extends to global health contexts; the framework requires adaptation in low- and middle-income settings where informatics infrastructure, governance capacity, and regulatory regimes differ substantially. Moreover, the objective structured clinical examination blueprint [44] and key performance indicators remain surrogate end points. By themselves, these measures do not guarantee improvements in patient-centered outcomes (eg, adverse events and readmissions). Second, a prospective, multicenter external evaluation is still necessary. Although we specify fairness analyses and minimum subgroup sample size and performance thresholds, real-world coverage across languages, cultural contexts, pediatrics, geriatrics, and rare-disease pathways is likely incomplete. Third, the regulatory environment remains dynamic as harmonized standards under the EU AI Act and FDA change control frameworks (eg, PCCPs) continue to evolve. Accordingly, operationalized procedures and performance thresholds should be periodically reassessed—particularly following material model updates—to sustain regulatory compliance and clinical relevance. Fourth, the automation paradox (skill decay) warrants attention. While AI agents improve efficiency, they may precipitate clinician deskilling over time. Safe fallback protocols (eg, reverting to manual workflows during system failure) are feasible only if clinicians maintain underlying diagnostic and procedural competence. To mitigate this risk, organizations should integrate automation-off scenarios into simulation training and downtime contingency plans to ensure clinicians remain capable of detecting errors and safely resuming control.

Given the rapid evolution of clinical AI, this framework requires ongoing refinement across 5 strategic areas. First, validation studies use expert consensus (eg, Delphi methods) and multicenter educational trials to link tiered competencies to observable clinical behaviors, such as error interception, safe deferral, and workflow efficiency. Second, assessment science strengthens psychometric measurement by refining objective structured clinical examination stations, bias and calibration checklists, and reliability targets (eg, generalizability coefficient [G] ≥0.70) [45,46]. Third, capacity building establishes faculty development programs and reusable educational resources—including deidentified sandboxes, annotated audit log exemplars, and HIPAA-aligned exercise sets—to support cross-specialty implementation. Fourth, advanced equity and uncertainty quantification cultivates practical competence in algorithmic fairness and uncertainty management through routine subgroup audits, coverage targets and reporting (using conformal prediction where appropriate), and bedside remediation playbooks. Fifth, simulation and institutional integration evaluate the feasibility and effectiveness of embedding automation-on and automation-off scenarios and rollback procedures within existing simulation programs and downtime contingency planning (eg, multidisciplinary team discussions). Outcomes include competency attainment, error interception during failures, auditability, and pathways for formal recognition within CME credit structures and institutional job descriptions.

## Conclusions

The progressive competency model integrates technical proficiency with ethical governance to provide clinicians with essential AI skills for the LLM era. By embedding these competencies into CME standards and job descriptions—using clear, observable behaviors—institutions can standardize safe and accountable AI use. Preparing for an AI-augmented future requires integrating governance-focused skills into medical training and professional development. This approach positions clinicians as responsible stewards of AI, ensuring adoption remains sustainable and centered on patient care.

### Acknowledgments

Following the initial drafting of the manuscript, the authors used ChatGPT (OpenAI) for artificial intelligence–assisted language editing to enhance clarity and style. All suggested edits were subject to critical review, and the authors subsequently made additional substantive revisions. The authors assume complete responsibility for the integrity and accuracy of the final content.

### Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

### Data Availability

Data sharing is not applicable to this paper as no new datasets were generated or analyzed during this study.

## Authors' Contributions

JL, QZ, and SL conceived and designed the study. WC and JL led data curation, with literature analysis by WC, JL, QZ, and SL. JL, WC, QZ, and SL contributed to writing the original draft. All authors critically revised the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

## Conflicts of Interest

None declared.

## References


1. Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, et al. JAMA Summit on AI. AI, health, and health care today and tomorrow: the JAMA Summit Report on Artificial Intelligence. JAMA. Nov 11, 2025;334(18):1650-1664. [doi: 10.1001/jama.2025.18490] [Medline: 41082366]
2. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. Mar 25, 2025. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2025-09-10]
3. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. Jun 28, 2023;25:e48568. [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
4. Liu S, Huang SS, McCoy AB, Wright AP, Horst S, Wright A. Optimizing order sets with a large language model-powered multiagent system. JAMA Netw Open. Sep 02, 2025;8(9):e2533277. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.33277] [Medline: 40986301]
5. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. Feb 05, 2025;388:e081554. [FREE Full text] [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
6. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. United States Food and Drug Administration. Aug 2025. URL: https://tinyurl.com/5dr3xx6j [accessed 2025-09-12]
7. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union. Jun 13, 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2025-09-10]
8. Augmented intelligence in health care. American Medical Association. 2018. URL: https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf [accessed 2025-09-16]
9. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. Jun 28, 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2025-09-10]
10. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The rise and potential of large language model based agents: a survey. Sci China Inf Sci. Jan 17, 2025;68(2):121101. [doi: 10.1007/s11432-024-4222-0]
11. Jiang Y, Black KC, Geng G, Park D, Zou J, Ng AY, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. NEJM AI. Aug 28, 2025;2(9):1-9. [doi: 10.1056/aidbp2500144]
12. Skalidis I, Maurizi N, Salihu A, Fournier S, Cook S, Iglesias JF, et al. Artificial intelligence and advanced digital health for hypertension: evolving tools for precision cardiovascular care. Medicina (Kaunas). Sep 04, 2025;61(9):1597. [FREE Full text] [doi: 10.3390/medicina61091597] [Medline: 41010987]
13. Liu J, Liu F, Fang J, Liu S. The application of Chat Generative Pre-trained Transformer in nursing education. Nurs Outlook. Nov 2023;71(6):102064. [doi: 10.1016/j.outlook.2023.102064] [Medline: 37879261]
14. Schuitmaker L, Drogt J, Benders M, Jongsma K. Physicians' required competencies in AI-assisted clinical settings: a systematic review. Br Med Bull. Jan 16, 2025;153(1):ldae025. [doi: 10.1093/bmb/ldae025] [Medline: 39821209]
15. Scott IA, Shaw T, Slade C, Wan TT, Barmanray R, Coorey C, et al. Proposing core competencies for physicians in using artificial intelligence tools in clinical practice. Intern Med J. Aug 2025;55(8):1403-1409. [doi: 10.1111/imj.70112] [Medline: 40576330]
16. Goodman KE, Rodman AM, Morgan DJ. Preparing physicians for the clinical algorithm era. N Engl J Med. Aug 10, 2023;389(6):483-487. [doi: 10.1056/NEJMp2304839] [Medline: 37548320]
17. Russell RG, Lovett Novak L, Patel M, Garvey KV, Craig KJ, Jackson GP, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med. Mar 01, 2023;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
18. Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. J Med Internet Res. Sep 15, 2025;27:e72644. [FREE Full text] [doi: 10.2196/72644] [Medline: 40955776]
19. Naveen K, Sajja WS, Nerella A. Building secure AI agents for autonomous data access in compliance/regulatory-critical environments. Comput Fraud Secur. 2024;2024(9):363-373. [FREE Full text] [doi: 10.52710/cfs.746]
20. Huang K. Agentic AI: Theories and Practices. Cham, Switzerland. Springer; 2025.

21. Acharya DB, Kuppan K, Divya B. Agentic AI: autonomous intelligence for complex goals—a comprehensive survey. IEEE Access. 2025;13:18912-18936. [doi: 10.1109/access.2025.3532853]
22. Borkowski AA, Ben-Ari A. Multiagent AI systems in health care: envisioning next-generation intelligence. Fed Pract. May 2025;42(5):188-194. [doi: 10.12788/fp.0589] [Medline: 40831649]
23. Lu JG, Song LL, Zhang LD. Cultural tendencies in generative AI. Nat Hum Behav. Nov 20, 2025;9(11):2360-2369. [doi: 10.1038/s41562-025-02242-1] [Medline: 40542181]
24. Grazhdanski G, Vasilev V, Vassileva S, Taskov D, Antova I, Koychev I, et al. SynthMedic: utilizing large language models for synthetic discharge summary generation, correction and validation. J Biomed Inform. Oct 2025;170:104906. [FREE Full text] [doi: 10.1016/j.jbi.2025.104906] [Medline: 40962129]
25. Jonnagaddala J, Wong ZS. Privacy preserving strategies for electronic health records in the era of large language models. NPJ Digit Med. Jan 16, 2025;8(1):1-3. [FREE Full text] [doi: 10.1038/s41746-025-01429-0] [Medline: 39820020]
26. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res. Aug 11, 2023;25:e48009. [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
27. Ganta T, Kia A, Parchure P, Wang MH, Besculides M, Mazumdar M, et al. Fairness in predicting cancer mortality across racial subgroups. JAMA Netw Open. Jul 01, 2024;7(7):e2421290. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.21290] [Medline: 38985468]
28. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health. Feb 26, 2025;24(1):57. [FREE Full text] [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
29. Mesinovic M, Watkinson P, Zhu T. Explainability in the age of large language models for healthcare. Commun Eng. Jul 17, 2025;4(1):128. [FREE Full text] [doi: 10.1038/s44172-025-00453-y] [Medline: 40676176]
30. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature. Jun 09, 2025;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
31. Berzin TM, Topol EJ. Preserving clinical skills in the age of AI assistance. The Lancet. Oct 2025;406(10513):1719. [doi: 10.1016/s0140-6736(25)02075-6]
32. Kim SH, Wihl J, Schramm S, Berberich C, Rosenkranz E, Schmitzer L, et al. Human-AI collaboration in large language model-assisted brain MRI differential diagnosis: a usability study. Eur Radiol. Sep 07, 2025;35(9):5252-5263. [doi: 10.1007/s00330-025-11484-6] [Medline: 40055233]
33. Dorfner FJ, Dada A, Busch F, Makowski MR, Han T, Truhn D, et al. Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. J Am Med Inform Assoc. Jun 01, 2025;32(6):1015-1024. [doi: 10.1093/jamia/ocaf045] [Medline: 40190132]
34. Makhni S, Rico J, Cerrato P, Hill B, Overgaard S, Wu J, et al. A comprehensive approach to responsible AI development and deployment. Mayo Clin Proc Digit Health. Dec 2025;3(4):100294. [FREE Full text] [doi: 10.1016/j.mcpdig.2025.100294] [Medline: 41282583]
35. M S AR, C R N, B R S, Lahza H, Lahza HF. A survey on detecting healthcare concept drift in AI/ML models from a finance perspective. Front Artif Intell. Apr 17, 2023;5:955314. [FREE Full text] [doi: 10.3389/frai.2022.955314] [Medline: 37139355]
36. Kore A, Abbasi Bavil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. Nat Commun. Feb 29, 2024;15(1):1887. [FREE Full text] [doi: 10.1038/s41467-024-46142-w] [Medline: 38424096]
37. Moskalenko V, Kharchenko V. Resilience-aware MLOps for AI-based medical diagnostic system. Front Public Health. Mar 27, 2024;12:1342937. [FREE Full text] [doi: 10.3389/fpubh.2024.1342937] [Medline: 38601490]
38. Augmented intelligence development, deployment, and use in health care. American Medical Association. Nov 2024. URL: https://www.ama-assn.org/system/files/ama-ai-principles.pdf [accessed 2025-09-16]
39. Augmented intelligence in medicine. American Medical Association. URL: https://www.ama-assn.org/practice-management/digital-health/augmented-intelligence-medicine [accessed 2025-09-10]
40. Digital education for building health workforce capacity. World Health Organization. Apr 14, 2020. URL: https://www.who.int/publications/i/item/9789240000476 [accessed 2025-09-10]
41. Global competency and outcomes framework for the essential public health functions. World Health Organization. Jun 10, 2024. URL: https://www.who.int/publications/i/item/9789240091214 [accessed 2025-09-18]
42. Artificial intelligence competencies for medical educators. Association of American Medical Colleges. URL: https://tinyurl.com/44t88c4c [accessed 2025-10-21]
43. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the health sector: systematic review of key skills for future health professionals. JMIR Med Educ. Feb 05, 2025;11:e58161. [FREE Full text] [doi: 10.2196/58161] [Medline: 39912237]
44. Milan FB, Grochowalski JH. A resource efficient and reliable standard setting method for OSCEs: borderline regression method using standardized patients as sole raters in clinical case encounters with medical students. Med Teach. Aug 2022;44(8):878-885. [doi: 10.1080/0142159X.2022.2041586] [Medline: 35234562]
45. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach. 2012;34(11):960-992. [doi: 10.3109/0142159X.2012.703791] [Medline: 23140303]

46. Park SY, Lee SH, Kim MJ, Ji KH, Ryu JH. Acceptability of the 8-case objective structured clinical examination of medical students in Korea using generalizability theory: a reliability study. J Educ Eval Health Prof. Sep 08, 2022;19:26. [FREE Full text] [doi: 10.3352/jeehp.2022.19.26] [Medline: 36071557]

## Abbreviations

**AI:** artificial intelligence
**CME:** continuing medical education
**EHR:** electronic health record
**EU:** European Union
**FDA:** Food and Drug Administration
**HIPAA:** Health Insurance Portability and Accountability Act
**LLM:** large language model
**PCCP:** predetermined change control plan
**WHO:** World Health Organization

*Edited by A Coristine; submitted 26.Oct.2025; peer-reviewed by S Palmieri, A Algumaei; comments to author 24.Nov.2025; revised version received 23.Dec.2025; accepted 23.Dec.2025; published 14.Jan.2026*

*Please cite as:*
*Cao W, Zhang Q, Liu J, Liu S*
*From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era*
*J Med Internet Res 2026;28:e86550*
*URL: https://www.jmir.org/2026/1/e86550*
*doi: 10.2196/86550*
*PMID:*