Large Language Models in Medicine: From Fluent Output to Trustworthy Clinical Implementation

Payam Hosseinzadeh Kasani [1, 2], Kee Hyun Cho [1, 2], Cheol-Heui Yun [3, 4, 5*]

[1] Department of Pediatrics, Kangwon National University Hospital, Chuncheon, Republic of Korea. [2] Department of Pediatrics, Kangwon National University School of Medicine, Chuncheon, Republic of Korea. [3] Department of Agricultural Biotechnology, and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea. [4] Center for Food and Bioconvergence, and Interdisciplinary Programs in Agricultural Genomics, Seoul National University, Seoul, Republic of Korea. [5] Institutes of Green Bio Science and Technology, Seoul National University, Pyeongchang, Republic of Korea.

Abstract

Large language models are rapidly entering medicine for documentation, patient communication, education, administration, research synthesis, and decision support, but their fluent outputs can be mistaken for trustworthy medical knowledge.

This review examines the limitations of large language models in medicine across three domains: first, epistemic reliability, which includes hallucinations, inconsistent outputs, outdated knowledge, and poor uncertainty calibration; second, clinical translation, which includes workflow disruption, diagnostic risk, unsafe patient-facing use, and overreliance; and third, governance accountability, which includes transparency, bias, privacy, cybersecurity, liability, and regulation.

Although LLMs may support selected clinical tasks and diagnostic second opinions, their context-dependent performance and hidden failure modes require local validation, human oversight, privacy protection, subgroup testing, auditability, and continuous governance to ensure they remain assistive rather than autonomous clinical tools.

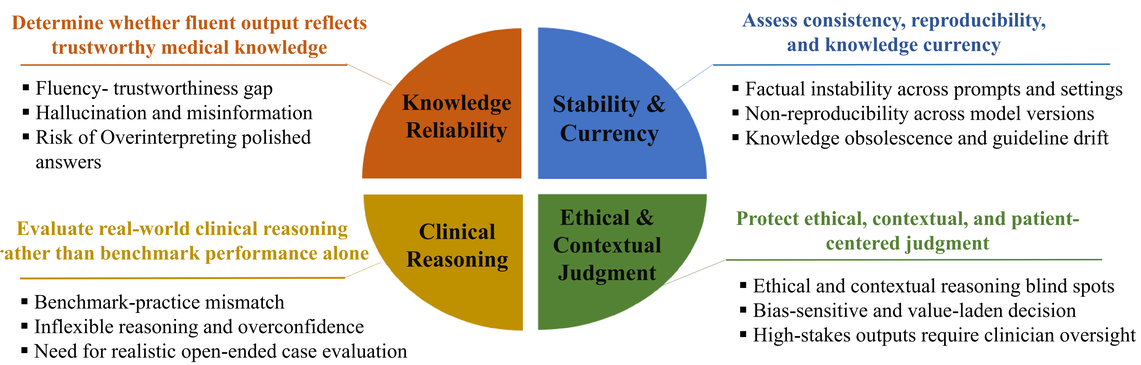

Figure 1. Epistemic reliability limitations of large language models in medicine

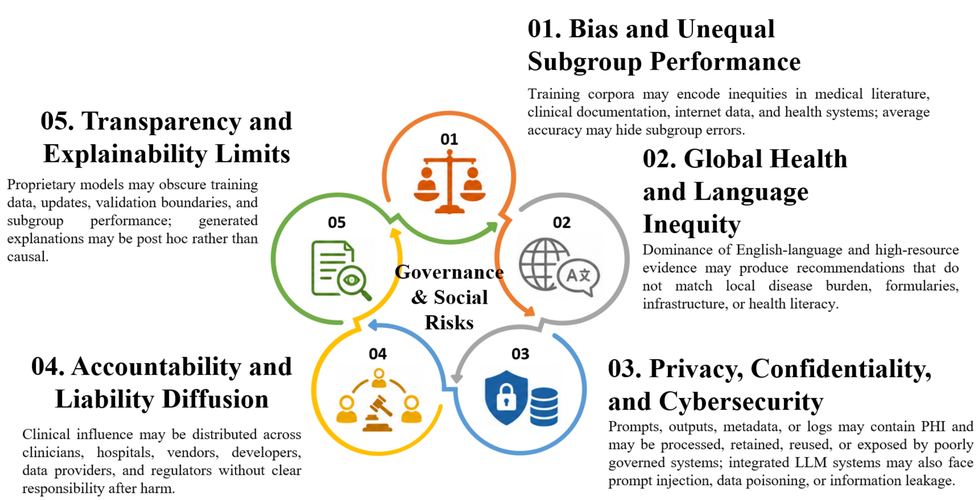

Figure 2. Trust at the System Level: Governance Risks in Medical LLM Deployment

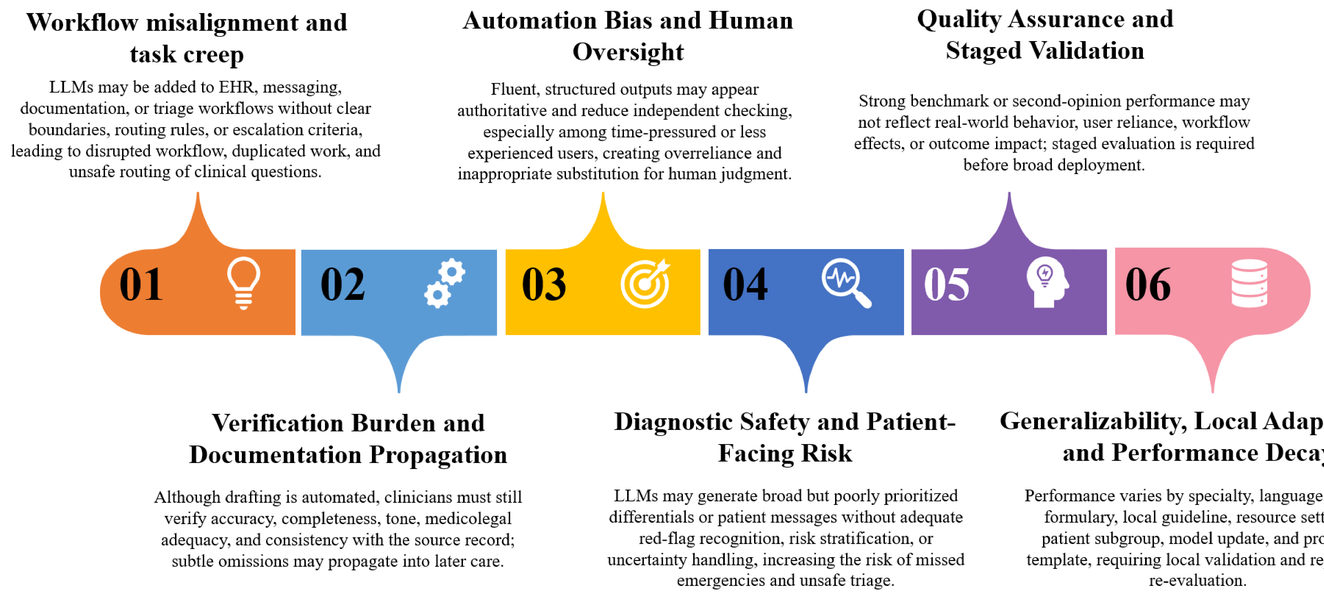

Figure 3. From Model Performance to Safe Practice: Major Clinical Translation Challenges for Medical LLMs

Table 1. Risk-based governance architecture for medical LLMs

Governance level	Core oversight functions	Required documentation or evidence
Use-case risk classification	Classify each application by intended use, autonomy, patient-facing status, clinical consequence, and user expertise.	Risk tier, intended-use statement, workflow map, prohibited uses, and approval decision.
Developer and vendor oversight	Assess safety, bias, privacy, cybersecurity, data handling, update policy, performance claims, and incident reporting obligations.	Model documentation, validation summary, security report, data-processing agreement, update log, and known limitations.
Institutional implementation	Approve local use cases, integrate with EHR safely, define prompts/templates, train users, and control access.	Local validation report, user-training records, prompt/version registry, privacy-impact assessment, and access-control plan.
Clinical human oversight	Verify outputs, escalate uncertainty, avoid unsupported high-risk use, maintain human judgment, and disclose AI involvement when appropriate.	Clinician review record, escalation pathway, patient-facing disclosure policy, and incident documentation.
Continuous monitoring and AI-QI	Monitor performance decay, subgroup errors, harmful omissions, drift, user behavior, and adverse events; revalidate after updates.	Audit dashboards, periodic performance reports, revalidation schedule, incident-review log, and stopping/modification criteria.
Regulatory, ethics, and publication oversight	Clarify medical-device or clinical decision-support status, define validation standards, require reporting, and support post-market surveillance.	Regulatory classification, ethics checklist, protocol approval, model/prompt disclosure, safeguards, and post-market monitoring plan.

Conclusion

The central challenge of medical LLMs is not their ability to produce fluent text, but whether that text can be trusted, safely translated into clinical workflows, and governed with clear accountability. This review therefore frames LLM deployment as a sociotechnical problem spanning epistemic reliability, clinical implementation, patient safety, equity, privacy, cybersecurity, transparency, and regulation. Ultimately, LLMs should be adopted only as assistive tools under local validation, human oversight, continuous monitoring, and institution-level governance, so that innovation improves care without weakening clinical responsibility or public trust