Position: Generalist Medical AI Requires Community-Governed Infrastructure

Anonymous Author(s)
Affiliation
Address
email

Abstract

This position paper argues that Generalist Medical AI (GMAI) cannot be achieved through model scaling alone, because the key bottleneck is not model capability but the absence of shared infrastructure for coordinating distributed medical intelligence. High-value medical data, clinical expertise, validation signals, and deployment feedback are decentralized, privacy-sensitive, and institutionally governed; this makes centralized approaches structurally insufficient. We argue that GMAI should be developed as open, community-governed, expert-endorsed infrastructure rather than as isolated proprietary models or fragmented specialist systems. We propose four necessary infrastructure mechanisms: resource discovery, living expert validation, governed deployment feedback, and sustainable community governance. Concrete instantiations include Doctor's Last Exam (a living expert-endorsed benchmark), PatientHub (shared longitudinal patient cases), Medical MCP (governed tool protocols), Agentic Hospital (deployment environments), and AI-native Health Information Systems. This position reframes GMAI as a socio-technical infrastructure agenda for trustworthy, scalable, and equitable medical intelligence.

1 Introduction

Recent progress in foundation models [Bommasani et al., 2021] has encouraged a scaling-centered view of Generalist Medical AI [Moor et al., 2023]: *larger models, broader modalities, and more tasks* may eventually yield systems that support diagnosis, triage, imaging, documentation, and care coordination. This view is powerful but incomplete.

Definition 1 *Generalist Medical AI is defined as a class of medical AI systems capable of generalizing across tasks, modalities, and clinical scenarios.*

GMAI Challenges Medicine differs from many general AI domains because its most valuable resources are structurally distributed. Medical data is highly sensitive, deeply professional, strongly regulated, and closely tied to patient privacy [Price and Cohen, 2019]. Much of the most valuable medical data remains distributed across hospital servers, research institutions, public health systems, and local clinical databases [Johnson et al., 2023]. These "sleeping data" and institutional silos make it unrealistic for any single company or centralized platform to collect, standardize, and govern all the data required for GMAI [Rieke et al., 2020]. Similarly, medical expertise is decentralized across specialties, institutions, regions, and clinical communities, making it difficult for proprietary systems to internalize all necessary knowledge in a scalable way. The development of GMAI therefore follows two complementary routes: 1) top-down scaling by large organizations and 2) bottom-up specialization by clinical communities. We argue that GMAI could be achieved through a *generalist-specialist integration* approach where generalist models provide broad multi-modal perception and high-level reasoning, while specialist components provide depth and precision [Nori et al., 2023b].

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

Figure 1: The challenge of sleeping medical data: high-value medical data and expertise are decentralized across hospitals, research institutions, public health systems, and clinical communities, making centralized collection unrealistic.

The key challenge is connecting these distributed components into a unified, trustworthy, and scalable ecosystem.

Our Position. We argue that such integration requires a community-driven mechanism. The central position of this paper is:

> **GMAI will fail if treated as a model-scaling race; instead, it should be built as open, community-governed, expert-endorsed infrastructure that coordinates fragmented medical resources and expertise.**

Community-driven GMAI is both a technical necessity and a governance response. Technically, trustworthy GMAI requires high-quality data, expert annotation, clinical feedback, rigorous benchmarking [Jin et al., 2021, Pal et al., 2022, Wang et al., 2024a], continuous evaluation, evidence-grounded retrieval [Lewis et al., 2020, Zakka et al., 2024], and human-in-the-loop deployment—all of which depend on distributed expertise and community participation. From a governance perspective, without open infrastructure, medical intelligence may become concentrated in proprietary systems whose incentives may not align with global public health needs [Reddy et al., 2020, World Health Organization, 2021]. An open, community-driven infrastructure can ensure that the benefits of medical AI are shared broadly rather than limited to a few companies, countries, or privileged groups.

2 Related Work

From Specialized to Generalist Medical AI Early medical AI systems were typically built for narrowly defined tasks, such as skin cancer classification [Esteva et al., 2017], chest X-ray diagnosis [Rajpurkar et al., 2017], and breast cancer screening [McKinney et al., 2020]. Despite strong performance in controlled settings, each system often required task-specific data, models, workflows, and deployment pipelines. Foundation models have shifted attention toward Generalist Medical AI (GMAI), where one system can operate across tasks, modalities, and clinical contexts [Moor et al., 2023]. Representative efforts include Med-PaLM and Med-PaLM 2 for medical question answering and licensing exams [Singhal et al., 2023, 2025], Med-Gemini for multimodal medical reasoning [Saab et al., 2024], and studies of general-purpose models such as GPT-4 in medical settings [Nori et al., 2023a]. Open medical models, including HuatuoGPT, PMC-LLaMA, LLaVA-Med, and BiomedGPT, further explore domain adaptation and multimodal understanding [Zhang et al., 2023a, Wu et al., 2024, Li et al., 2023, Zhang et al., 2023b]. In contrast, we argue that GMAI is not only a model-scaling problem, but also an infrastructure problem: it requires an open ecosystem that connects general models, specialist tools, clinical knowledge bases, and local workflows.

Medical Data, Knowledge, and Evaluation Infrastructure A core barrier to medical AI is that clinical data remains distributed across hospitals, research institutions, and public health systems.

Federated learning offers one path to use such data without centralization [Rieke et al., 2020, Dayan et al., 2021], while public datasets such as MIMIC-IV [Johnson et al., 2023] and resources such as UMLS [Bodenreider, 2004] support research but capture only part of real-world medicine. Retrieval-augmented generation links models to external evidence [Lewis et al., 2020], and systems such as Almanac suggest that guideline grounding can improve clinical reliability [Zakka et al., 2024, Xiong et al., 2024]. Evaluation has also advanced through benchmarks such as MedQA, PubMedQA, MedMCQA, CMB, and MedBench [Jin et al., 2021, 2019, Pal et al., 2022, Hendrycks et al., 2021, Wang et al., 2024a, Cai et al., 2024, Tu et al., 2024]. Yet many benchmarks remain static and exam-oriented, falling short on clinical judgment, safety, evidence use, and workflow fit. We therefore view data, knowledge, and evaluation as shared infrastructure that requires expert input, provenance tracking, and community maintenance.

Medical Agents and Open Governance Recent work explores medical AI in agentic and workflow-based settings. Agent Hospital simulates hospital roles and workflows for training medical agents [Li et al., 2024], AgentClinic evaluates agents in simulated clinical encounters [Schmidgall et al., 2024], and MedAgents explores multi-agent collaboration [Tang et al., 2024]. These works show that medical AI must interact with tools, records, and institutional workflows beyond text generation. Healthcare AI also raises governance questions around accountability, transparency, safety, and equity [Reddy et al., 2020, World Health Organization, 2021]. We extend these discussions by proposing open, community-driven GMAI infrastructure with shared licenses, expert-endorsed resources, community benchmarks, feedback channels, and sustainable institutions that align medical AI with public health needs rather than proprietary interests alone.

Gap in Existing Work. While these works advance medical AI capabilities, evaluation methods, and agentic systems, a critical gap remains: *no existing work proposes integrated infrastructure for coordinating distributed medical resources, expert validation, and governed deployment*. Federated learning addresses data privacy but not benchmarking, knowledge bases, or feedback loops. Open models provide technical artifacts but not governance mechanisms or quality assurance. Medical agents demonstrate workflow capabilities but lack real-world deployment infrastructure. Our framework fills this gap by proposing community-driven infrastructure that connects distributed data access, living expert validation, governed deployment feedback, and sustainable governance into a unified ecosystem for trustworthy GMAI.

3 Why GMAI needs Community-driven Infrastructure, Not Only Models

3.1 Why Top-down Scaling Alone is Insufficient

The development of medical AI has followed two complementary routes. The first is a top-down scaling route, where large organizations build powerful foundation models using advantages in data, compute, and infrastructure [Singhal et al., 2023, Saab et al., 2024, Nori et al., 2023a]. The second is a bottom-up specialization route, where researchers, clinicians, startups, and hospitals build models, datasets, tools, and workflows for specific diseases, specialties, departments, or care settings [Zhang et al., 2023a, Chen et al., 2023, Wu et al., 2024, Li et al., 2023]. In many general AI domains, top-down scaling can absorb large amounts of public or easily standardized data. In medicine, however, this assumption breaks down.

Challenges of Top-down Scaling in GMAI Top-down scaling assumes that large organizations can accumulate enough data, compute, and infrastructure to build broadly capable medical models. This assumption breaks down in medicine: high-value data is sensitive, regulated, institutionally governed, and often remains "sleeping" across hospitals, research institutions, public health systems, specialist centers, and regional healthcare networks [Price and Cohen, 2019, Johnson et al., 2023]. Such data are hard to centralize due to privacy, incentives, incompatible formats, consent, and local governance constraints [Rieke et al., 2020, Dayan et al., 2021]. Expertise is likewise distributed across specialties, regions, and care roles. Centralized efforts such as Watson Health, and the gap between exam-benchmark performance and clinical reliability in frontier models, suggest that the bottleneck is not model capability alone, but distributed data access and expert validation.

General-specialist Integration for GMAI. The practical roadmap of GMAI is unlikely to be purely general or purely specialized. Generalist models can provide broad reasoning, multimodal perception, natural language interaction, and task coordination. Specialist models, disease-specific knowledge bases, clinical tools, local protocols, and expert systems provide depth, precision, and

Table 1: Comparison of approaches to medical AI development. Community-driven GMAI infrastructure could address limitations of both centralized proprietary systems and open models alone.

Aspect	Centralized Proprietary	Open Models Only	Community-driven GMAI
Data access	Limited to partnerships	Public datasets only	Federated + governed registries
Expertise	Internal teams	Volunteer contributors	Credentialed experts + incentives
Evaluation	Internal validation	Static benchmarks	Living expert-endorsed evaluation
Deployment	Controlled rollout	Uncontrolled release	Risk-stratified governed access
Governance	Corporate decisions	No formal governance	Community-endorsed mechanisms
Sustainability	Profit-driven	Volunteer-dependent	Reciprocal value creation

adaptation [Nori et al., 2023b]. The challenge is not whether generalist or specialist systems will prevail, but how to connect them into a trustworthy, scalable, and governable infrastructure.

3.2 Reasons for Community-driven Infrastructure

General-specialist integration cannot be achieved by model composition alone. It requires a shared infrastructure that coordinates generalist models, specialist resources, clinical tools, expert knowledge, benchmarks, and deployment feedback. Such infrastructure is not merely a collection of open-source models or volunteer contributions, but a set of governed mechanisms through which distributed medical resources can be discovered, validated, connected, and improved over time.

Therefore, GMAI requires more than larger models, open releases, or isolated specialist systems. It requires community-driven infrastructure with three core necessities. This infrastructure is also a governance response: without open and accountable mechanisms, medical intelligence may become concentrated in a few proprietary platforms whose incentives may not align with public health needs [Reddy et al., 2020, World Health Organization, 2021]. Notably, even proprietary organizations have increasingly recognized the need for community-like expert participation, as shown by HealthBench [Arora et al., 2025] and HealthBench Professional [Hicks et al., 2026], which draw on the expertise of hundreds of physicians across multiple countries.

Necessity I: Leveraging Segmented Resources Without mechanisms for accessing distributed medical resources, GMAI cannot move beyond narrow public datasets or proprietary hospital partnerships. Medical data and expertise are segmented across hospitals, specialties, regions, public health systems, and patient communities, and cannot be fully centralized because of privacy, consent, institutional governance, and local clinical context. Community infrastructure should therefore provide metadata registries, provenance tracking, privacy-preserving access, expert directories, and shared access rules, so that locally governed resources become visible without being extracted into a single platform.

Necessity II: Expert-endorsed Living Validation Medical AI cannot be trusted through static benchmarks or automatic metrics alone. Trustworthy GMAI requires expert annotation, clinical judgment, safety review, evidence grounding, and continuous benchmark updates [Jin et al., 2021, Pal et al., 2022, Wang et al., 2024a, Tu et al., 2024]. Community-driven evaluation turns benchmarking from a one-time leaderboard into a living validation process, where expert users continuously identify unsafe responses, missing evidence, workflow failures, and region-specific needs. This makes evaluation part of system improvement rather than a final reporting step.

Necessity III: Governed Deployment Feedback GMAI must improve from real-world use, but clinical deployment feedback cannot be collected through uncontrolled logging or purely commercial extraction. Hospitals, doctors, patients, medical schools, and startups need reciprocal mechanisms: contributors provide cases, annotations, tools, or workflow feedback, and in return receive better models, safer assistants, reusable benchmarks, and shared infrastructure. Governed feedback channels, responsible licenses, audit trails, and risk-aware access can transform deployment from a one-way extraction process into an accountable learning loop aligned with safety, sustainability, and equity [Reddy et al., 2020, World Health Organization, 2021].

3.3 Addressing Common Objections

This position paper addresses four common objections.

Objection 1: "Proprietary systems are safer because medical AI is high-risk." Closed development may provide control, but control is not the same as safety. Medical AI safety depends

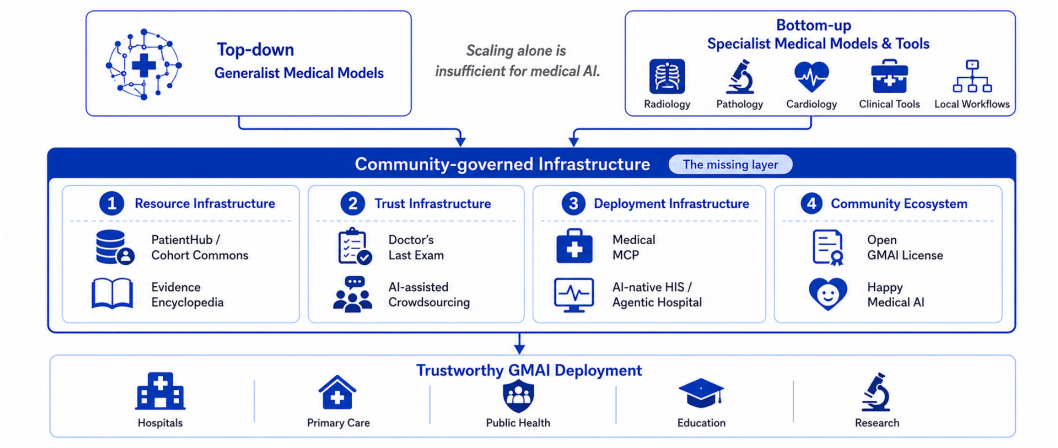

Figure 2: The Framework of GMAI.

on transparent evaluation, independent expert review, audit trails, evidence versioning, and post-deployment monitoring [Reddy et al., 2020, World Health Organization, 2021]. These mechanisms can be strengthened, not weakened, by community infrastructure. Openness should therefore mean governed transparency, not uncontrolled release.

Objection 2: "Federated learning already solves decentralized medical data." Federated learning is important, but it solves only part of the problem [Rieke et al., 2020, Dayan et al., 2021]. GMAI also needs shared benchmarks, expert consensus, clinical knowledge bases, tool standards, workflow environments, and continuous feedback. Federated learning might be a component of community-driven GMAI, not a substitute for it.

Objection 3: "Specialist models are enough; we do not need GMAI." Specialist models are essential, but clinical practice is rarely confined to one specialty. Patients often have multiple diseases, long histories, medications, social contexts, and operational constraints. The goal of GMAI is not to replace specialists, but to coordinate them through general-specialist integration: broad reasoning from generalist models combined with depth from specialist models, tools, and knowledge bases [Nori et al., 2023b].

Objection 4: "Open medical AI increases misuse risk." This concern is valid, but the answer is not naive openness or complete closure. We advocate *governed openness*: credential-verified contribution, risk-stratified access, responsible licensing, audit logging, and human approval for high-risk actions. Such mechanisms make open medical AI more accountable while still allowing the community to inspect, improve, and govern shared infrastructure.

4 Framework for Community-Driven GMAI

Community-driven Generalist Medical AI should not be understood as a single foundation model released to the public. It is a shared infrastructure for turning distributed medical data, expert knowledge, evaluation signals, deployment feedback, and governance capacity into reusable medical intelligence. We organize this infrastructure into four interdependent pillars: (1) **Resource Infrastructure** for models, data, cohort resources, and expert knowledge; (2) **Trust Infrastructure** for benchmarks, crowdsourced annotation, evidence validation, and feedback loops; (3) **Deployment Infrastructure** for tools, agentic environments, agent platforms, and AI-native healthcare systems; and (4) **Community Ecosystem** for education, entrepreneurship, governance, multilingual participation, and long-term maintenance. These pillars jointly provide a practical route toward trustworthy, scalable, and globally accessible GMAI.

4.1 Resource Infrastructure: Models and Data

The first pillar provides foundational resources: open models, shared data, and expert knowledge. These components are tightly coupled: models require data and knowledge, while data and knowledge require models to become useful.

4.1.1 Open-sourced Models

Open-sourced models [Wu et al., 2024, Zhang et al., 2023a, Chen et al., 2023, Touvron et al., 2023] are necessary but insufficient. Medical AI involves patient privacy, institutional consent, clinical safety, and deployment accountability [World Health Organization, 2021, Reddy et al., 2020]. A community-driven ecosystem requires governance specifying what can be shared, by whom, under what conditions, and with what obligations. Moreover, open models must support deployment across heterogeneous hardware ecosystems including NVIDIA, AMD or Ascend accelerators; so that GMAI infrastructure is not locked to a single vendor and can be deployed in regions with different hardware availability and regulatory requirements.

Open GMAI License A community-developed Open GMAI License can define how datasets, models, tools, benchmarks, and knowledge bases are shared, reused, and deployed. Community funding can support open-source organizations, annotation fellowships, hospital pilots, and global programs, transforming "sleeping medical data" into public-interest infrastructure.

4.1.2 Shared Data

Rather than centralizing sensitive data, the community can build a registry where institutions publish metadata, provenance, access rules, and consent conditions. Actual data remains locally stored or accessed through privacy-preserving computation, while the registry records what exists, who contributed it, and how it may be used.

PatientHub: Longitudinal Cases and Cohort Commons. A representative example is PatientHub, a governed infrastructure for longitudinal patient cases and cohort-level resources. Each case records a patient's care trajectory across symptoms, visits, diagnoses, laboratory tests, imaging, medications, interventions, outcomes, and follow-up. These cases can be organized into cohorts by disease, phenotype, treatment pathway, care setting, demographic profile, or outcome, supporting model training, evaluation, medical education, comparative analysis, and workflow simulation. PatientHub therefore turns individual care trajectories into reusable cohort resources while preserving privacy through de-identification, synthetic or hybrid case construction, provenance tracking, consent documentation, and clear intended-use rules.

4.2 Trust Infrastructure: Benchmarking and Crowdsourcing

Trust infrastructure ensures GMAI resources meet clinical standards through rigorous evaluation and expert validation, combining benchmarking and crowdsourced annotation.

4.2.1 Benchmark

Medical evaluation requires expert judgment and real clinical context beyond static benchmarks or automatic metrics. A trustworthy infrastructure combines benchmark construction, expert review, post-deployment feedback, and continuous updates.

Doctor's Last Exam. Inspired by expert-level benchmarks designed to evaluate frontier AI capabilities across broad academic domains, a crowd-funded "Doctor's Last Exam" becomes a living examination program jointly maintained by clinicians, medical schools, and hospitals. Cases cover diagnosis, triage, longitudinal care, multimodal interpretation, and workflow execution. Doctors correct unsafe responses, patients report usability issues, and hospitals provide workflow feedback, creating a loop from real-world needs to benchmark design and system improvement.

4.2.2 Crowdsourcing

Medical annotations require clinical training, specialty knowledge, and awareness of uncertainty. A platform should verify credentials, route tasks to suitable experts, and record annotator background and confidence. It supports dataset construction, benchmark authoring, RAG validation, and feedback collection.

AI-Assisted Crowdsourcing Platform. AI systems draft labels, extract findings, and propose diagnoses. Verified experts review, correct, or approve outputs, focusing on high-value judgment rather than repetitive screening. The platform supports audit trails, privacy review, disagreement

resolution, and quality control. This creates a loop where AI prepares annotations, experts provide judgment, corrected labels improve models, and improved systems assist future annotation.

AI-native Evidence-based Encyclopedia-like Guidelines. An AI-native medical encyclopedia should be community-owned, organized by specialty, and governed by expert-endorsed editorial structures. Each claim carries provenance, evidence level, update history, and clinical scope. The encyclopedia supports retrieval, citation, version control, and API access, enabling GMAI to rely on current evidence rather than static memory. Research access remains free, while contributor incentives tie to usage records. Revenue from clinical or educational use returns to authors and the community.

4.3 Deployment Infrastructure: Agentic Tools and Environments

The third pillar enables GMAI to move from research prototypes to deployable systems. This layer provides the technical infrastructure for medical agents to interact with tools, hospital systems, and real-world workflows.

4.3.1 Medical Tool (MCP)

GMAI requires a shared medical tool layer where community-validated APIs, calculators, and workflow interfaces can be safely used by agents. Models should call tools for current evidence, calculations, safety checks, or human approval. Standardized protocols enable tools from hospitals, specialty groups, and startups to be discovered, invoked, and audited under common rules.

Medical MCP A community-based medical API hub documents each tool with its schema, provenance, validation status, risk level, and required permission. Low-risk tools may be called automatically; high-risk tools require explicit confirmation or expert review. This allows agents to combine general reasoning with trustworthy utilities while making invocations traceable and governable.

4.3.2 Agentic Environments

GMAI should be deployed as agentic systems capable of routing tasks, invoking tools, retrieving knowledge, and coordinating workflows [Li et al., 2024, Schmidgall et al., 2024]. This requires environments with shared tools, standardized patient simulation, and HIS interfaces to train, evaluate, and safely deploy agents before clinical use.

Agentic Hospital An agentic hospital environment simulates hospital roles, departments, patient trajectories, devices, and HIS interactions. It supports training, evaluation, safety testing, and workflow design through reusable standardized patient simulators. At deployment, this layer connects agents to AI-native HIS infrastructure with secure, permission-controlled, auditable interfaces and routing mechanisms [Tang et al., 2024].

AI-native HIS AI-native Health Information Systems (HIS) differs from Agentic Hospital environments. Agentic Hospital provides simulated organizational settings for training and testing; AI-native HIS provides agent-compatible software infrastructure for interacting with healthcare data and workflows. Unlike traditional systems designed for human interfaces, AI-native HIS supports structured APIs, tool-calling protocols, access control, audit logging, and human approval. Together, they create the operating environment for testable, deployable agentic GMAI.

Agent Platform A community GMAI portal allows doctors to test copilots, hospitals to evaluate workflow agents, students to train with standardized patients, and researchers to access benchmarks. It collects real demands: where models fail, what evidence is missing, which workflows are hard to automate. For willing contributors, it provides governed channels for sharing de-identified cases, feedback logs, and annotations under explicit consent and privacy review. The platform becomes both an access point and a feedback engine for improving open medical intelligence.

4.4 Community Ecosystem: Education, Entrepreneurship, and Governance

The fourth pillar ensures long-term viability of the GMAI ecosystem. This addresses community building, economic sustainability, and institutional governance that transforms GMAI into a durable public infrastructure.

4.4.1 Medical Education

GMAI requires an open medical AI education ecosystem. Inspired by communities like Datawhale, the community can build a "Happy Medical AI" initiative: a free, practice-oriented curriculum for doctors, students, patients, developers, and public health workers to understand, use, evaluate, and contribute to medical AI. This layer combines online courses, tutorials, benchmark challenges, and contribution portals with offline bootcamps, hospital workshops, annotation camps, and hackathons to train contributors, collect needs, validate tools, and build trust.

Online Tutorial (Happy Medical AI). Inspired by community-driven open education projects, Happy Medical AI provides modular courses, clinical case exercises, annotation practice, benchmark tasks, agent-building examples, and deployment playbooks. Different tracks serve different communities: clinicians learn evidence retrieval and workflow support; students practice reasoning with standardized patients; developers learn data governance, tool APIs, and evaluation; patients learn health literacy and safe AI use. Education becomes an entry point for participation, turning learners into contributors.

4.4.2 Open Entrepreneurship and Community Governance

Beyond research and technical infrastructure, GMAI requires sustainable governance mechanisms.

Open Entrepreneurship. Dedicated medical AI seed funds and community-based startup programs can help early-stage teams build on shared GMAI infrastructure. Potential directions include specialty agents, AI-native HIS tools, patient-facing applications, education platforms, workflow automation, multilingual services, annotation platforms, and privacy-preserving deployment. The goal is to create a healthy ecosystem where open infrastructure enables responsible innovation while startups contribute improvements, tools, and feedback back to the commons.

Global Community and Multilingual Support. A truly community-driven GMAI must be globally inclusive. This requires multilingual model support [Wang et al., 2024b, Zheng et al., 2024], localized medical guidelines, cross-cultural clinical knowledge, and regional community chapters. The infrastructure should support major languages (English, Chinese, Spanish, Arabic, Hindi, etc.) and enable local communities to contribute region-specific medical knowledge, traditional medicine practices, and culturally appropriate healthcare workflows. Global coordination mechanisms (including regional hubs, international workshops, and cross-border collaboration platforms) can connect medical professionals, researchers, and developers worldwide, ensuring that GMAI serves diverse populations rather than replicating existing healthcare inequalities.

5 Call for Actions in GMAI

5.1 What We Should Do Now

Community-driven GMAI will not emerge spontaneously. It requires deliberate action from specific stakeholders. We outline concrete next steps.

ML researchers should prioritize building shared evaluation infrastructure over publishing yet another medical benchmark paper. Contribute cases to Doctor's Last Exam, develop evaluation protocols that test clinical safety and workflow fitness (not just accuracy on multiple-choice questions), and release evaluation code and data under open licenses. When developing medical AI models, support deployment across heterogeneous hardware—including both NVIDIA GPUs and domestic accelerators such as Huawei Ascend—to avoid vendor lock-in.

Hospitals and health systems should participate in data registries that publish metadata, provenance, and access conditions without centralizing sensitive data. Provide deployment feedback on where AI systems fail in real workflows. Contribute de-identified cases to shared repositories under governed consent frameworks. The return is access to better models, benchmarks, and tools built on the collective infrastructure.

Industry should adopt open licensing frameworks (Open GMAI License) that define responsible sharing, reuse, and deployment obligations. Contribute tools to Medical MCP with documented schemas and validation status. Fund open infrastructure as shared pre-competitive investment rather than treating all medical AI as proprietary advantage.

Medical schools and professional societies should integrate AI literacy into medical education, contribute expert annotations through credentialed platforms, and participate in benchmark governance. Medical expertise is the scarcest resource in GMAI development; professional communities must organize to provide it sustainably.

Policymakers should develop regulatory frameworks that enable cross-border medical data governance, recognize community-endorsed evaluation as a pathway to clinical validation, and fund open medical AI infrastructure as public health investment. Current regulations often assume centralized development; policy must evolve to support distributed, community-driven approaches.

5.2 Applications

GMAI infrastructure enables diverse applications across healthcare settings. In hospitals, it supports clinical workflows, imaging interpretation, and evidence-grounded decision support [Topol, 2019, Tu et al., 2024]. For patients, it enables continuous longitudinal care, chronic disease management, and access to medical intelligence in underserved regions [Thirunavukarasu et al., 2023]. In public health, it facilitates population-scale disease surveillance and emergency response [Dayan et al., 2021, Rieke et al., 2020]. For education and research, it provides scalable learning infrastructure and accelerates biomedical innovation [Schmidgall et al., 2024, Jin et al., 2021]. Finally, it can serve as the reasoning layer for embodied healthcare agents including surgical and rehabilitation robots [Li et al., 2024]. This broad application landscape demonstrates why GMAI must be community-driven: no single institution can cover all healthcare scenarios, specialties, workflows, and populations.

6 Challenges and Limitations

We acknowledge significant challenges as below that must be addressed for GMAI vision to succeed.

Scope of This Work. As a position paper, our proposed framework and programs are forward-looking and have not yet been fully implemented or empirically validated. The concrete programs (e.g., Doctor's Last Exam and PatientHub) represent aspirational designs whose feasibility depends on sustained community participation, institutional buy-in, and adequate funding. We present this framework as a research agenda and call to action rather than a validated solution.

Coordination Costs and Governance Complexity. Open communities face inherent coordination overhead. Aligning diverse stakeholders (including hospitals, researchers, clinicians, patients, regulators, and industry) across different countries, legal systems, and incentive structures is substantially harder than centralized development. Decision-making may be slow, and governance structures may struggle to balance inclusivity with efficiency.

Quality Control and Safety Risks. Open contributions introduce quality variance. Unlike proprietary systems with unified quality pipelines, community-contributed data, annotations, and tools may vary in reliability. In healthcare, quality failures carry patient safety risks. Our framework addresses this through expert endorsement, credential verification, and tiered access controls, but these mechanisms add friction and may not catch all errors.

Privacy and Regulatory Heterogeneity. Medical data governance varies dramatically across jurisdictions. GDPR, HIPAA, China's Personal Information Protection Law, and other frameworks impose different requirements on data sharing, consent, and cross-border transfer. Building truly global infrastructure while respecting all regulatory regimes is an unsolved challenge that may limit the scope of data sharing in practice.

7 Conclusion

Generalist Medical AI should not be pursued through model scaling alone. Because medical data, expertise, trust, and deployment feedback are inherently distributed, GMAI requires an open, community-driven, and expert-endorsed infrastructure. This paper argues for building such infrastructure around shared resources, trustworthy evaluation, agentic deployment, and sustainable governance. By connecting generalist models with specialist knowledge, clinical feedback, responsible licensing, and community participation, community-driven GMAI offers a practical path toward scalable, trustworthy, and equitable medical intelligence for global healthcare.

References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arber, Sydney von Arx, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. MedBench: A large-scale Chinese benchmark for evaluating medical large language models. In *AAAI*, 2024.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. HuatuoGPT-II, one-stage training for medical adaption of LLMs. In *arXiv preprint arXiv:2311.09774*, 2023.

Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27(10):1735–1743, 2021.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

Rebecca Soskin Hicks, Mikhail Trofimov, Dominick Lim, Rahul K Arora, Foivos Tsimpourlas, Preston Bowman, Michael Sharman, Chi Tong, Kavin Karthik, Arnav Dugar, et al. Healthbench professional: Evaluating large language models on real clinician chats. *arXiv preprint arXiv:2604.27470*, 2026.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *EMNLP*, 2019.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*, 2023.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023a.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023b.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning (CHIL)*, 2022.

W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature Medicine*, 25 (1):37–43, 2019.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P Lungren, and Andrew Y Ng. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Sandeep Reddy, Sonia Allan, Simon Coghlan, and Paul Cooper. A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(3): 491–497, 2020.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3:119, 2020.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, et al. Capabilities of Gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31 (3):943–950, 2025.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of ACL*, 2024.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8): 1930–1940, 2023.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, et al. Towards generalist biomedical AI. *NEJM AI*, 1(3), 2024.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. CMB: A comprehensive medical benchmark in Chinese. In *NAACL*, 2024a.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, et al. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*, 2024b.

World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Technical report, World Health Organization, 2021.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. PMC-LLaMA: Toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of ACL*, 2024.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2), 2024.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. HuatuoGPT, towards taming language model to be a doctor. In *Findings of EMNLP*, 2023a.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Xun Chen, et al. BiomedGPT: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023b.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts. *arXiv preprint arXiv:2410.10626*, 2024.