



Large language models require a new form of oversight: capability-based monitoring

Katherine C. Kellogg, Bingyang Ye, Yifan Hu, Guergana K. Savova, Byron Wallace & Danielle S. Bitterman



Large language models (LLMs) have been rapidly adopted in healthcare, but oversight strategies are lacking. We propose capability-based monitoring, motivated by the fact that LLMs are generalist systems whose overlapping internal capabilities are reused across numerous downstream tasks. This approach organizes monitoring around shared capabilities to enable cross-task detection of systemic weaknesses, long-tail errors, and emergent behaviors. We describe considerations for developers, organizational leaders, and professional societies, and policymakers.

The enthusiasm and rapid uptake of generalist artificial intelligence (AI) models, in particular, large language models (LLMs), in healthcare have spurred much discussion of their evaluation and oversight for clinical applications. However, less attention has been paid to the core assumptions about model performance degradation that underpin monitoring strategies, but that break down in the case of LLM use.

To address this, we propose a new capability-based monitoring framework that is better aligned with how LLMs are trained and used in practice. Capability-based monitoring is an organizing principle for generalist LLM oversight that tracks shared model functions across downstream tasks such as summarization or information retrieval, enabling cross-task detection of systemic weaknesses, long-tail errors, and emergent behaviors. We detail (1) a preliminary taxonomy of LLM capabilities, (2) monitoring dimensions and potential metrics, (3) implementation and governance implications of adopting this framework, and (4) future research directions to fully realize the potential of capability-based monitoring. We provide technical recommendations for methodologists and researchers alongside organizational considerations for healthcare leaders and policymakers.

Why task-based monitoring fails for LLMs

Traditionally, AI implementations in healthcare have been focused on bespoke Machine Learning (ML) models, each trained for a single task using datasets from defined, bounded populations (Fig. 1, ML Paradigm). These models assume that training and test data come from the same underlying distributions. When this assumption is violated, overfitting occurs, leading to degraded performance on new datasets^{1,2}. ML models trained for a particular task, such as sepsis prediction³, on bounded, labeled clinical datasets reflecting (hopefully) their target clinical population, will always be overfit

(that is, only performant for the task and populations they were trained on) to some extent^{3,4}. Because of this, performance degradation of ML models post-deployment is a given: models will always degrade because populations and outcome distributions inevitably change compared to the data the model was trained on^{1,2}. This has led, sensibly, to model-specific post-deployment monitoring for expected degradation.

In contrast, the emergence of generalist LLMs fundamentally challenges these prior assumptions, driving performance monitoring. Despite not being trained using in-distribution clinical data or specifically for clinically-relevant tasks, LLMs can still capably summarize clinic visit transcripts into note drafts (ambient documentation)^{5,6}, answer clinical questions⁷, translate patient instructions⁸, and more. Inevitable model degradation due to changes in populations compared to the training dataset cannot be assumed, because LLMs were not trained for any specific task in any given population (Fig. 1, LLM Paradigm). Many—probably most—clinical tasks will be “out-of-distribution” for the LLM training set, which is massive and often unknown anyway. Thus, traditional notions of ML overfitting and performance degradation do not straightforwardly apply to LLMs. Performance variation due to LLM “overfitting” now occurs due to changes in contextual factors such as the knowledge, culture, deployment environments, or prompting structures, rather than changes from the training dataset distributions defined by input features and labels. While we shouldn’t anticipate degradation due to overfitting in the traditional sense, we do expect that an LLM will behave differently across populations in ways that are not necessarily predictable.

The generalist properties of LLMs make them powerful and drive uptake, but also complicate monitoring. Accordingly, for LLMs and other similar generalist models, monitoring frameworks must evolve. Ongoing task-based oversight is not only impractical as LLMs drive task expansion, but also undesirable because it will leave us blind to shared vulnerabilities. We therefore propose a new organizing principle guiding generalist LLM monitoring that is grounded in how these models are developed and used in practice: capability-based monitoring.

What is capability-based monitoring?

Capability-based monitoring—monitoring models around shared model capabilities, such as summarization, reasoning, or translation, in order to enable cross-task detection of systemic weaknesses, long-tail errors, and emergent behaviors—is motivated by the fact that LLMs are generalist systems whose overlapping internal capabilities are reused across numerous downstream tasks (Table 1, Fig. 2). We define capabilities as general-purpose LLM functions such as summarization that are the core measurable building blocks for many different tasks. In this approach, capabilities are the unit of monitoring: Tasks relying on the same underlying model and drawing on similar capabilities are monitored collectively—acknowledging that some tasks engage multiple capabilities simultaneously. For instance,

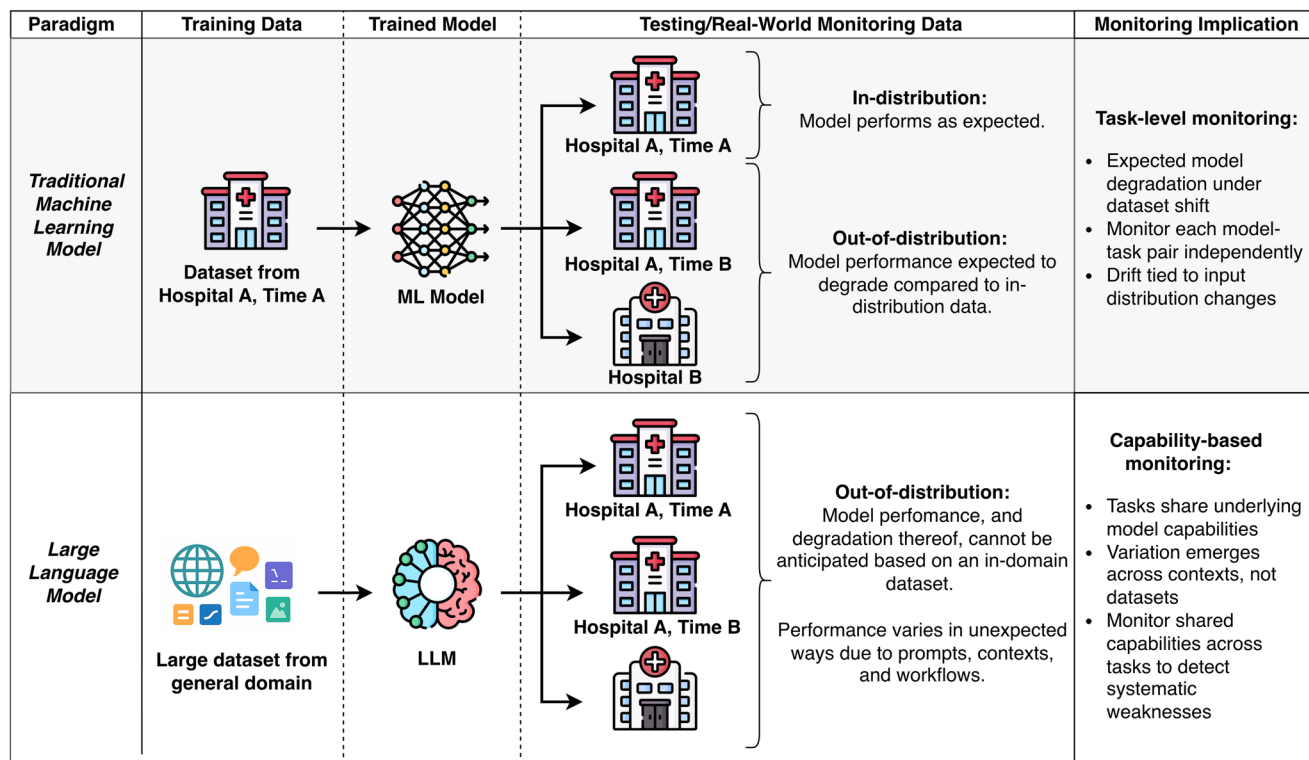


Fig. 1 | Illustration of train and test data distributions in traditional Machine Learning (ML) models vs. Large Language Models (LLMs). In traditional ML models, it is assumed that test data come from the same underlying distribution (i.e., in distribution; Hospital A, Time A in the figure). As models are applied to different real-world data distributions such as evolution over time (e.g., Hospital A, Time B)

and/or new settings (e.g., Hospital B), performance optimized and reported on in-distribution data is no longer reliable. Instead, performance is anticipated to degrade due to overfitting. In LLMs, models are trained from large, general datasets and learn general abilities. All clinical datasets are out of distribution, and traditional notions of ML overfitting and population drift do not straightforwardly apply.

the ability to summarize underlies a range of tasks with distinct contexts, such as inpatient discharge summary generation, outpatient pre-charting, and ambient documentation. Monitoring each task in isolation fragments oversight and risks missing cross-cutting vulnerabilities that propagate across tasks. In contrast, capability-based monitoring (Table 2) provides a more practical and comprehensive framework, enabling cross-task evaluation of shared operations, early detection of systemic weaknesses, and identification of edge cases or rare errors that task-specific monitoring might overlook (Fig. 2). This is particularly critical for LLMs, which often struggle with infrequent but clinically significant long-tail scenarios⁷.

The proposed framework: capabilities, dimensions, and metrics

LLM behavior depends on intrinsic and extrinsic contextual factors, and LLM performance degradation arises when a model “overfits” these contextual factors that shape its behavior. Intrinsic factors pertain to properties of the model itself, including its alignment with professional standards and values, temporal currency (i.e., how up-to-date its knowledge base is), reasoning quality, robustness to variation in input style or language, and computational efficiency. Extrinsic factors involve human interaction, including the degree of human oversight and the type and extent of human-model collaboration, both of which impact overall system performance⁹. When a model “overfits” these factors, it starts relying too heavily on specific contexts it has seen before, rather than maintaining adaptability. For example, real-world facts and language change over time. If

the model is tuned too tightly to outdated knowledge, it struggles with new information. Table 1 outlines a preliminary taxonomy of LLM capabilities, while Table 2 outlines monitoring dimensions (such as reasoning quality and input robustness) and proposed metrics for measuring these intrinsic and extrinsic factors.

Not all dimensions currently validated have automatable monitoring approaches that are known to correlate with human evaluation; many still require human review and gold-standard comparators, although LLM evaluation strategies and metrics, including gaps specific to healthcare, have been extensively discussed in prior work¹⁰⁻¹². Our framework aims to organize and prioritize metric development and validation. Existing benchmarks in both general and clinical domains, while imperfect, can also supplement real-world monitoring by identifying performance gaps within specific capabilities¹³.

Given the limited availability of validated metrics and ground truth labels they require, the LLM-as-judge paradigm (where a separate model is used to evaluate outputs) is gaining traction as a flexible, extensible monitoring method. We include LLM-as-judge as an automatable metric across several dimensions, but emphasize that these secondary models also require validation and ongoing oversight for each dimension in which they are applied (see Governance/Meta-Control Capability Family, Table 1).

A monitoring strategy should not only identify errors, but also lead to actionable corrections¹⁴. Importantly, performance degradation across dimensions may not always necessitate a full model update. Limitations arising from intrinsic factors may often be addressed through prompt

Table 1 | Generalist large language model capability taxonomy

Capability Family	Capability	Core function	Example tasks that require capability
Content Transformation: Capabilities focused on generating, modifying, or structuring raw data inputs	Summarization	Ability to compress text, preserve relevant facts, prioritize appropriate facts, and documents	Hospital course summary generation for discharge notes, ambient documentation, discharge instructions, insurance appeals
	Information Extraction	Identify and capture structured data from unstructured documentation Extract discrete fields, classification	Trial matching, medical coding, chart curation, medical coding
	Classification	Assign predefined codes or categories to standardized concepts	Patient portal triage, outcome prediction, medical coding
	Simplification	Generate patient-friendly language (lay and multi-lingual), reply to patients	Patient education, discharge instructions, patient question-answering, consent form simplification
	Translation	Multi-lingual translation and communication	Translation of patient-facing information, real-time translation in clinic
Clinical knowledge: Capabilities that use information to inform a user’s decision or answer a query	Clinical Reasoning	Given inputs, provide an appropriate clinical rationale or justification	Clinical decision support, insurance appeals, patient question-answering
	Information Retrieval	Identify and retrieve appropriate documents for an incoming query	Literature search, chart curation
Governance/Meta-Control: Capabilities that monitor system self-regulation, safety, and adherence to ethical or operational boundaries.	Factual and Clinical Alignment	Assess adherence to professional and institutional standards and policies; detection of fabricated, hallucinated, or outdated content	Auditing and fact-checking models, often described as LLM-as-judge. Note that these are meta-capabilities, and systems may be designed with embedded controls or separate guardrail models.
	Safety and Refusal Guardrails	Decide when to answer or block a response, safe completion, or refusal of unsafe/uncertain requests	
	Appropriate Language and Tone Monitoring	Detect and filter toxic, stigmatizing, or otherwise inappropriate language	

LLM Large Language Model

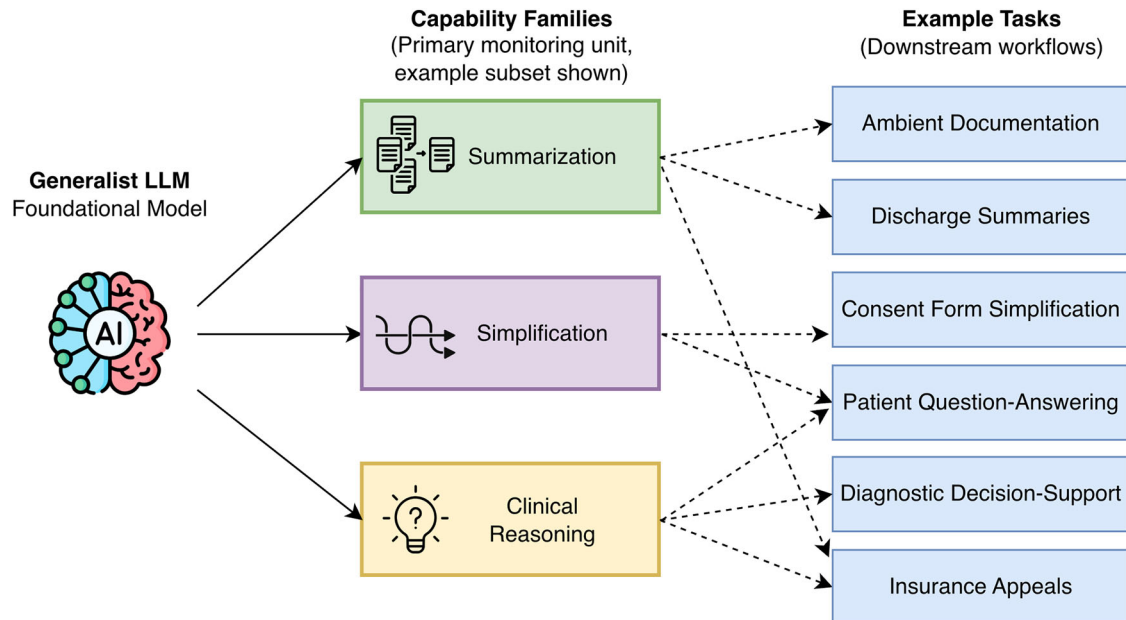


Fig. 2 | Conceptual monitoring stack for capability-based oversight of clinical large language model (LLM) systems. This conceptual stack organizes monitoring around three layers: a generalist LLM foundation model, intermediate capability families (the primary monitoring unit), and downstream clinical tasks/workflows.

Because multiple tasks draw on shared capabilities, monitoring signals can be aggregated at the capability layer across tasks rather than tracked separately for each task. Only an example subset of capability families and tasks is illustrated to demonstrate shared capabilities and instances of tasks using more than one capability.

Table 2 | Monitoring dimensions and proposed metrics

Monitoring dimension	Description	Example metrics that require human labels	Example automated metrics	Priority monitoring dimensions for each capability										
				Summarization	Information Extraction	Classification	Clinical Reasoning	Simplification	Translation	Information retrieval	Governance/Meta-Control			
Information quality and accuracy	Faithfulness to inputs, factual correctness	% factual errors in sampled outputs scored by a human expert	LLM-as-judge Medical knowledge benchmarks	X	X	X	X	X	X	X	X	X	X	X
Reasoning	Internal logic of output and clinical soundness	Human expert review	LLM-as-judge Medical reasoning benchmarks				X				X			
Style	Clarity, readability, and tone appropriateness	Human expert review	Automated readability metrics LLM-as-judge	X			X	X	X	X				
Sycophancy* and refusal behavior	Ability to resist declining unsafe or uncertain requests	Human expert review	LLM-as-judge Sycophancy benchmarks				X				X			X
Input robustness and feature drift	Stability under changing prompts, data quality	Human expert review	Overlap metrics LLM-as-judge Text-based statistics (e.g., input tokens)	X	X	X	X	X	X	X	X	X	X	X
Equity	Performance across subgroups (demographics, specialties)	N/A	Distribution of other metrics across subgroups, including raw distribution and fairness metrics	X	X	X	X	X	X	X	X	X	X	X
End-user preference	Human edit rates, acceptance ratios, and escalation frequency	N/A	Edit distance, acceptance rate	X	X	X	X	X	X	X	X	X	X	X
Toxicity	Presence of toxic, stigmatizing, or otherwise inappropriate language	Human expert review	LLM-as-judge Toxicity and bias benchmarks	X			X	X	X	X	X	X	X	X
Process	Costs, energy, time	N/A	Tokens, costs, FLOPs, latency per unit time (to understand usage), and per query (to understand potential LLM behavior changes)	X	X	X	X	X	X	X	X	X	X	X

LLM Large Language Model, FLOPS Floating point operations per second
 *Sycophancy is defined as excessive agreeableness, at the expense of accuracy and/or safety.

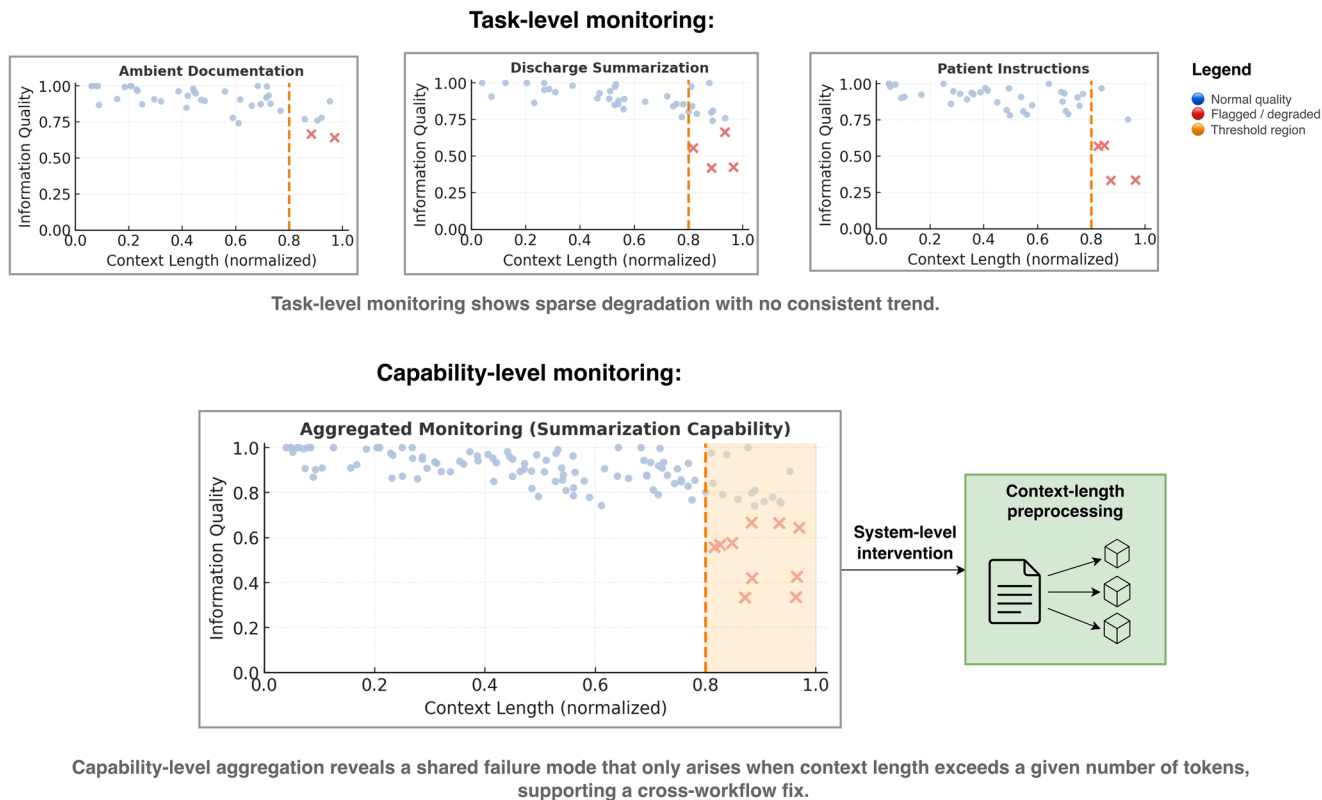


Fig. 3 | Aggregating task-level signals via capability-based monitoring reveals shared failure modes. Individual monitoring shows sparse quality issues across hospital course discharge summarization, ambient documentation, and patient

instructions. Aggregated capability-level monitoring exposes a shared vulnerability, for example, errors arising when context length exceeds a threshold, enabling corrective preprocessing that restores performance across all summarization workflows.

refinement, improved tool integration, or adjustments to retrieval databases before modifying the underlying model. In contrast, extrinsic factors may call for interventions such as enhanced interface design, user training, or targeted education. We envision primary capability monitoring occurring on a per-LLM basis, as each model is trained on distinct datasets that are typically opaque to the institutions deploying them. Nevertheless, given the shared pretraining corpora and similar tuning paradigms among many LLMs, and the fact that it may not always be known when a vendor updates LLM’s weights, vulnerabilities identified in one model should prompt systematic evaluation across related models.

As an example of the strengths of capability-based monitoring, envision an institution that has implemented 3 tasks requiring strong summarization capabilities: hospital course summarization, ambient documentation, and patient-facing discharge instructions (Fig. 3). Missing information is flagged sparsely in all 3 tasks, and the signal only becomes significant when grouped, enabling identification that errors occur when input exceeds a context length threshold, thus a solution is implementation of a new preprocessing step to reduce context length. Similarly, a rare token (wordpiece) repeated many times in a single patient’s inpatient notes is found to trigger biased language. Efficient simulations confirm this is a shared failure mode, so an input filter is implemented for all summarization tasks, avoiding future potential errors for all summarization workflows.

Implications for developers and methodologists

Implementing capability-based monitoring creates new challenges, implications, and benefits for the healthcare community (Table 3). Key challenges

for developers in healthcare organizations include: (a) capability and monitoring dimensions are not yet fully scoped and taxonomized, and will increase over time, (b) it is not feasible to manually monitor all of these metrics for all models, and (c) human oversight and task-specific monitoring will likely still be required for very high-risk applications. Developers can address these challenges by (a) developing visualization-based dashboards and defining evaluation frequency and thresholds for error detection, (b) engaging in two-tiered monitoring with automated screening by LLMs-as-judge and existing automated metrics (high-frequency, low-cost) and human review of flagged cases (low-frequency, high-interpretability), and (c) experimenting with various techniques for addressing human automation bias, over-reliance, and de-skilling in order to provide effective human oversight of models deployed in high-risk tasks.

Implications for institutions and policymakers

Key challenges for organizational leaders in healthcare organizations include: a) decentralized capability-based monitoring at the business unit-level gives business unit leaders more control, but risks missing cross-cutting vulnerabilities, b) merely detecting performance degradation through capability-based monitoring is not sufficient, c) individuals may develop LLM implementations via prompt refinement for private use and not report these to the organization for monitoring, and (d) use of LLMs can deskill healthcare workers, making it difficult to take LLMs offline when deterioration is detected. Organizational leaders can address these challenges by (a) centralizing capability-based monitoring while working with business unit stakeholders to identify and create specific data views and functionality

Table 3 | Monitoring implementation and oversight design

New Monitoring Challenges with LLMs		Implications for Practice	Benefits	Limitations	Specific Recommendations
Implications for Developers					
Task-based monitoring fragments oversight and misses cross-cutting vulnerabilities	<ul style="list-style-type: none"> Create registries/ dashboards that visualize performance metrics per capability, not per task 	<ul style="list-style-type: none"> Scalability and reduced redundancy in compliance and auditing 	<ul style="list-style-type: none"> Capability and monitoring dimensions not yet fully scoped and taxonomized, and will increase over time 	<ul style="list-style-type: none"> Develop visualizations of capability-based dashboards Define evaluation frequency and thresholds for error detection Audit and log LLM use in a standardized fashion, e.g., using MedLog¹⁵, extended to include capability family/families for the task 	
Not feasible to manually monitor all of these metrics for all models	<ul style="list-style-type: none"> Implement existing automatic metrics and identify gaps therein Develop new automated metrics based on identified gaps New automated metrics may include Judge LLMs: generative models used to evaluate outputs of other LLMs 	<ul style="list-style-type: none"> Scalable, continuous, and low-cost oversight 	<ul style="list-style-type: none"> Need to "audit the auditor" via periodic human calibration 	<ul style="list-style-type: none"> Recommend tiered framework: <ul style="list-style-type: none"> Automated screening by Judge LLMs and existing automated metrics (high-frequency, low-cost) Human review of flagged cases (low-frequency, high-interpretability) For smaller, more resource-constrained organizations, a "lite" tier with core safety guardrails and summarization monitoring using simple sampling plus manual review may be sufficient 	
Some truly high-risk implementations will merit their own individualized oversight (e.g., models making treatment recommendations without a human-in-the-loop)	<ul style="list-style-type: none"> Maintain risk-stratified evaluation of an emerging technology. High-risk devices still need the appropriate clinical testing before being integrated and monitored For models that are integrated, there will be a risk threshold at which organizations decide they still need individualized monitoring, but that will be the minority of cases 	<ul style="list-style-type: none"> Human oversight of very high-risk models 	<ul style="list-style-type: none"> Humans may miss errors due to automation bias, over-reliance, and de-skilling 	<ul style="list-style-type: none"> Work with clinicians to investigate the feasibility of lim-as-judge or other monitoring method for a tiered approach over time 	
Difficult to support the needs of diverse stakeholders (health system leaders, clinical experts, and technical personnel who are distributed across the organization) with a standardized set of metrics	<ul style="list-style-type: none"> Identify key stakeholder groups; conduct participatory design sessions with diverse stakeholders to develop prototypes of monitoring dashboards 	<ul style="list-style-type: none"> Supports teams of health system leaders, clinical experts, and technical personnel that are distributed across the organization as they monitor and respond to model deterioration 	<ul style="list-style-type: none"> Varied levels of technical expertise and knowledge may limit communication and understanding of metrics While streamlined compared to task-based monitoring, rapidly expanding capabilities may require ongoing reassessments 	<ul style="list-style-type: none"> Identify specific data, views and functionality required by different stakeholders Periodically re-evaluate monitoring needs with stakeholders as models advance 	
Identified performance degradation will need to be addressed	<ul style="list-style-type: none"> Develop a standardized approach for root cause analysis Develop methods for correcting LLM performance Create back-up strategies for critical LLM-mediated functions 	<ul style="list-style-type: none"> Enables a resilient model ecosystem that is robust to failures LLM performance degradation will not always require model fine-tuning; rapid prompt engineering and agentic updates may solve the problem. Capability-based monitoring enables shared solutions across workflows 	<ul style="list-style-type: none"> Limited insight and control over vendor LLMs Increasingly complex agentic systems with tool use and retrieval complicate root cause identification and resolution The same fix may not always work for all tasks, increasing workload 	<ul style="list-style-type: none"> Create best practices for manual review of errors, prompt review, and agentic system review Maintain ongoing communication and collaboration with vendors Ensure failure is due to LLM itself and not the surrounding architecture, which may be less generalizable Establish and maintain a database of example inputs for all workflows to confirm shared failure mode and resolution Maintain shared database of errors and solutions 	

Table 3 (continued) | Monitoring implementation and oversight design

New Monitoring Challenges with LLMs		Implications for Practice	Benefits	Limitations	Specific Recommendations
Implications for organizational leaders					
Capability-based monitoring at the business unit-level provides control for business-unit leaders, but risks missing cross-cutting vulnerabilities	<ul style="list-style-type: none"> Centralize capability-based monitoring 	<ul style="list-style-type: none"> Centralized monitoring by capability reduces monitoring burden across hundreds of use cases, and enables cross-context evaluation of shared operations, early detection of systemic weaknesses, and identification of edge cases or rare errors 	<ul style="list-style-type: none"> Centralization reduces customization of solutions for each business unit and reduces the overall responsiveness to business unit needs Capability-based monitoring appropriate for post-deployment monitoring is not a substitute for initial needs assessment and evaluation 	<ul style="list-style-type: none"> Build a team and resources to centralize capability-based monitoring Identify specific data views and dashboard functionality required by business unit stakeholders. Continue to perform the initial needs assessment and evaluation by model 	
Merely detecting performance degradation is not sufficient	<ul style="list-style-type: none"> Identify who is accountable for diagnosing the root cause of degradation and applying strategies to restore model performance Develop a set of methods for root cause diagnosis and for restoring model performance Identify who needs to be informed of model issues, including taking models offline 	<ul style="list-style-type: none"> Ensures that degradations in model performance will be addressed, and the estimated ROI will continue to be realized 	<ul style="list-style-type: none"> Limited insight and control over vendor LLMs Increasingly complex agentic systems with tool use and retrieval complicate root cause identification and resolution The same fix may not always work for all tasks, increasing the workload 	<ul style="list-style-type: none"> Maintain ongoing communication and collaboration with vendors Review failures with business unit leaders to ensure a comprehensive understanding of failures and fixes Establish collaborations with other institutions to share identified errors and resolutions 	
Use of LLMs can deskill healthcare workers, making it difficult to take LLMs offline when deterioration is detected	<ul style="list-style-type: none"> Institute requirements that professionals regularly practice high-impact tasks without AI, to maintain proficiency Leverage simulation technology to maintain task proficiency 	<ul style="list-style-type: none"> Can enable early detection of AI-induced deskillings in high-expertise domains 	<ul style="list-style-type: none"> Tradeoffs between deskillings solutions that minimize deskillings and those that impose additional time and effort demands 	<ul style="list-style-type: none"> Institute requirements that professionals regularly practice mission-critical tasks without AI, to maintain proficiency 	
Lack of clear regulations makes it difficult to determine monitoring metrics	<ul style="list-style-type: none"> Integrate with regulatory and accreditation processes by partnering with government affairs teams to create awareness of government agencies/regulatory bodies, and iteratively developing governance policies 	<ul style="list-style-type: none"> Supports the iterative development of metrics based on changing regulations 	<ul style="list-style-type: none"> AI technology will continue to move faster than external regulations 	<ul style="list-style-type: none"> Craft internal governance principles and governance process in advance of regulations Continue to monitor external regulations to align internal processes with new regulations 	
Speed of change in models makes it difficult to determine which capabilities should be monitored	<ul style="list-style-type: none"> Assign responsibility for external environmental scanning for new model capabilities 	<ul style="list-style-type: none"> Supports the monitoring of new capabilities Related capabilities may help anticipate future needs and failure modes Potential for cross-institution collaboration to learn from others' experiences 	<ul style="list-style-type: none"> Automated evaluation metrics will lag behind capability emergence, requiring more intensive initial manual oversight Emerging capabilities may initially resemble more traditional task-based, single workstream monitoring, which may require bespoke visualizations and metrics Potential increased computational resource requirements for new models and monitoring thereof create a bottleneck limited by the institutional infrastructure and/or cost 	<ul style="list-style-type: none"> Establish internal team to review the literature for new capabilities, monitoring methods, and solutions Institute best practices for integrating a new capability family into the monitoring dashboard Create a communication structure for developers, informatics, and clinical team members to report gaps in capabilities Maintain reporting pathways for ad hoc error detection and requirements for critical harm reporting Create a strategy for prioritizing model assessments to manage computational/cost resources 	
Individuals may develop LLM implementations for private use and not report these to the organization for monitoring	<ul style="list-style-type: none"> Develop pathways and incentives for reporting bespoke workflows to the organization 	<ul style="list-style-type: none"> Supports monitoring of all models being used by organization members 	<ul style="list-style-type: none"> Ease of developing new LLM workflows complicates the identification and tracking of all uses Need for additional resources to identify and integrate uses into centralized monitoring systems 	<ul style="list-style-type: none"> Reward formalization of models: Provide recognition and rewards for formalizing new models Increase benefits associated with formalization: Provide resources for integration of models into the EMR system so they can be part of everyday workflow 	

Table 3 (continued) | Monitoring implementation and oversight design

New Monitoring Challenges with LLMs	Implications for Practice	Benefits	Limitations	Specific Recommendations
<p>Comprehensive monitoring introduces privacy and disclosure considerations, particularly for event-level logs</p>	<p>• Create governance standards for the monitoring framework itself, and integrate the system within established security infrastructure</p>	<p>• Centralized and consistent monitoring enables standardized data privacy practices and broad disclosure</p>	<p>• Privacy risks and disclosure needs inherently scale with monitoring, although this is not unique to capability-based monitoring</p> <p>• Event-level logs are useful for auditing, troubleshooting, and accountability, but can contain sensitive information</p>	<p>• Engage legal and ethical bodies, as well as patient stakeholders, to establish institutional standards for disclosure/consent and education around comprehensive monitoring</p> <p>• Include monitoring systems in primary AI governance review</p> <p>• Engage information security teams for unified security/review and integration into the data ecosystem</p>
<p>Implications for professional associations and policymakers</p>				
<p>Different sophistication in LLM monitoring across institutions</p>	<p>• Develop shared benchmarks and reference frameworks across institutions</p>	<p>• Supports unified registries, clearer accountability, and consistent safety reporting</p>	<p>• In-house technical expertise required</p> <p>• LLM-extrinsic monitoring dimensions may be highly sensitive and unique to specific institutions</p>	<p>• Encourage sharing of frameworks, benchmarking strategies, and other monitoring resources via publication, presentation, and funding opportunities</p> <p>• Formalize working groups and special conferences/workshops for dissemination and training</p>
<p>Inconsistent safety reporting</p>	<p>• Collaborative “monitoring commons” for healthcare AI safety</p>	<p>• Supports unified registries, clearer accountability, and consistent safety reporting</p>	<p>• Institutions must commit to logging LLM use according to shared protocols and taxonomies</p>	<p>• Centralized nationwide database for reporting LLM issues</p>

LLM Large Language Model, ROI Return on Investment.

required by these decentralized stakeholders, (b) identifying who is accountable for diagnosing the root cause of degradation, and developing a set of methods for root cause diagnosis and restoring model performance, (c) providing recognition, rewards and resources to individuals for formalizing new models, and (d) instituting requirements that professionals regularly practice high-impact tasks without AI, to maintain proficiency.

Addressing these challenges will pay dividends in the future for institutions and regulators. Capability-based monitoring aligns with the EU AI Act’s post-market monitoring requirements for AI-enabled medical devices, which necessitate procedures to track data, assess high-risk system performance, and identify and address emerging risks¹⁵. Similarly, the US FDA’s lifecycle management approach to AI monitoring will likely incorporate ongoing real-world monitoring, including protocols to identify model and data drift^{16,17}.

By measuring consistent metrics and dimensions, capability-based monitoring will support regulators in identifying performance drift and rare error modes at scale. Regulators, payers, and industry should support shared monitoring infrastructure, standardized logging formats, and regional collaboratives, which would be particularly beneficial for resource-constrained settings.

Limitations and future research directions

While capability-based monitoring should enable more practical and robust oversight, future work is needed to realize its full potential at scale.

First, our proposed capabilities and metrics are likely not exhaustive, and we encourage the community to contribute to a comprehensive taxonomy of each. As discussed above, there is an ongoing need for development and validation of automated performance metrics relevant to healthcare applications¹⁰⁻¹². Although more streamlined than task-based monitoring, there are still many ways an organization may wish to visualize capability across models and business units. Research into the optimal visualizations and interface for such monitoring tools will be needed to make sure they are usable and sustainable. Evaluation frequency and thresholds for error detection across all monitoring dimensions will need to be defined and refined as we gain experience implementing LLMs. The quality assurance, process improvement, and statistical quality control fields will play an important role in developing these thresholds.

Second, the monitoring infrastructure itself will have a high cost and resource burden, posing challenges for resource-constrained institutions and institutions without existing informatics infrastructure. Implementing multi-dimensional monitoring across numerous capabilities, particularly when relying on LLM-as-judge models, requires substantial, specialized computational resources and personnel that may be prohibitive. Furthermore, scalability is limited by dependence on human labeling for ground truth to validate and audit metrics, which poses a constraint on the rate at which new dimensions can be validated and implemented. Further, extensive monitoring itself introduces expanded needs for data privacy and disclosure/consent. Regulator, payer, and industry support for shared monitoring would offset burdens.

Because many institutions will use the same underlying LLMs for various tasks, there is an enormous opportunity to extend this strategy to a collaborative monitoring commons across institutions. While capability-based collaborative monitoring will not require sharing data or models, it will require standardized documentation and logging of LLM use. Active uptake and expansion of efforts such as MedLog, a protocol for event-level clinical AI logging, to include capability tagging and associated metadata, will be critical in realizing this vision¹⁸.

Conclusions

In the LLM era, “overfitting” in healthcare AI has shifted from model training to prompt, context, and workflow over-adaptation, making the traditional distinction between in-distribution and out-of-distribution clinical data far less predictive of performance. Monitoring of generalist AI, exemplified by LLMs, should be fit-for-purpose: designed to address how LLMs are trained and used in practice, not simply extended from traditional models that have different performance and generalization assumptions. As such, the unit of monitoring must evolve from tasks to capabilities, tracking shared behaviors across contexts. Capability-based monitoring is at once technically necessary and organizationally scalable. Healthcare systems, vendors, and regulators should adopt capability-based frameworks to ensure safe, equitable, and sustainable deployment of generalist AI.

Data availability

No datasets were generated or analyzed during the current study.

Katherine C. Kellogg¹ ✉, Bingyang Ye^{2,3,4}, Yifan Hu⁵,
Guergana K. Savova⁶, Byron Wallace⁷ & Danielle S. Bitterman^{2,3,6} ✉

¹MIT Sloan School of Management, Boston, MA, USA. ²AI in Medicine Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA. ³Department of Radiation Oncology, Brigham and Women’s Hospital/Dana-Farber Cancer Institute, Boston, MA, USA. ⁴Department of Computer Science, Brandeis University, Waltham, MA, USA. ⁵Harvard John A. Paulson School Of Engineering And Applied Sciences, Cambridge, MA, USA. ⁶Computation Health Informatics Program, Boston Children’s Hospital, Harvard Medical School, Boston, MA, USA. ⁷Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA.

✉e-mail: kkellogg@mit.edu; dbitterman@bwh.harvard.edu

Received: 1 December 2025; Accepted: 1 May 2026;
Published online: 15 May 2026

References

- Wong, A. & Sussman, J. B. Understanding model drift and its impact on health care policy. *JAMA Health Forum* **6**, e252724 (2025).
- Finlayson, S. G. et al. The Clinician and Dataset Shift in Artificial Intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).
- Wong, A. et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* **181**, 1065–1070 (2021).
- Habib, A. R., Lin, A. L. & Grant, R. W. The epic sepsis model falls short—the importance of external validation. *JAMA Intern. Med.* **181**, 1040–1041 (2021).
- Afshar, M. et al. A Novel Playbook for Pragmatic Trial Operations to Monitor and Evaluate Ambient Artificial Intelligence in Clinical Practice. *NEJM AI* **2**, <https://doi.org/10.1056/aidbp2401267> (2025).
- You, J. G. et al. Ambient documentation technology in clinician experience of documentation burden and burnout. *JAMA Netw. Open* **8**, e2528056 (2025).
- Chen, S. et al. The effect of using a large language model to respond to patient messages. *Lancet Digit. Health* **6**, e379–e381 (2024).
- Ray, M. et al. Evaluating a large language model in translating patient instructions to Spanish using a standardized framework. *JAMA Pediatr.* **179**, 1026–1033 (2025).
- Sujan, M. et al. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform.* **26**, e100081 (2019).
- Agrawal, M., Chen, I. Y., Gulamali, F. & Joshi, S. The evaluation illusion of large language models in medicine. *NPJ Digit. Med.* **8**, 600 (2025).
- Moreno, A. C. & Bitterman, D. S. Toward clinical-grade evaluation of large language models. *Int. J. Radiat. Oncol. Biol. Phys.* **118**, 916–920 (2024).
- Sai, A. B., Mohankumar, A. K. & Khapra, M. M. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.* **55**, 1–39 (2023).

- Ni, S. et al. A survey on large language model benchmarks. *ArXiv abs/2508.15361*, (2025).
- Guan, H., Bates, D. & Zhou, L. Keeping Medical AI Healthy and Trustworthy: A Review of Detection and Correction Methods for System Degradation. In *IEEE Transactions on Biomedical Engineering*. <https://doi.org/10.1109/TBME.2025.3642706> (2025).
- EU Artificial Intelligence Act. <https://artificialintelligenceact.eu/>.
- Request For Public Comment: Measuring and Evaluating Artificial Intelligence-enabled Medical Device Performance in the Real-World. U.S. Food and Drug Administration https://www.fda.gov/medical-devices/digital-health-center-excellence/request-public-comment-measuring-and-evaluating-artificial-intelligence-enabled-medical-device?utm_medium=email&utm_source=govdelivery (2025).
- Warraich, H. J., Tazbaz, T. & Califf, R. M. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA* **333**, 241–247 (2025).
- Noori, A. et al. A global log for medical AI. *arXiv [cs.AI]* (2025).

Acknowledgements

The authors acknowledge financial support from the National Institutes of Health National Cancer Institute (U54CA274516-01A1, R01CA294033-01, U24CA248010, the American Cancer Society, American Society for Radiation Oncology, ASTRO-CSDG-24-1244514-01-CTPS Grant DOI #: <https://doi.org/10.53354/ACS.ASTRO-CSDG-24-1244514-01-CTPS.pc.gr.222210>, a Patient-Centered Outcomes Research Institute (PCORI) Project Program Award (ME-2024C2-37484), and the Woods Foundation. All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors, or Methodology Committee. This work was also conducted with support from UM1TR004408 award through Harvard Catalyst — the Harvard Clinical and Translational Science Center (National Center for Advancing Translational Sciences, National Institutes of Health) and financial contributions from Harvard University and its affiliated academic health-care centers.

Author contributions

K.K., B.Y., Y.H., G.S., B.W., and D.B. conceptualized the idea. K.K. and D.B. wrote the main manuscript text and prepared the tables. B.Y. and Y.H. prepared the figures. K.K., B.Y., Y.H., G.S., B.W., and D.B. edited the manuscript. K.K., B.Y., Y.H., G.S., B.W., and D.B. read and approved the final manuscript.

Competing interests

D.S.B.: Associate Editor, J.C.O. Clinical Cancer Informatics (not related to the submitted work), Associate Editor, Annals of Oncology (not related to the submitted work), Associate editor of Radiation Oncology of HemOnc.Org (not related to the submitted work) and is on the Scientific Advisory Board of Mercurial AI (not related to the submitted work) and Blue Clay Health LLC (not related to the submitted work). The authors declare no other competing interests.

Additional information

Correspondence and requests for materials should be addressed to Katherine C. Kellogg or Danielle S. Bitterman.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026