The Compute Allocation and Prioritization Brief for Hospital Leaders

A free guide to deciding which clinical AI workloads come first when computing capacity is tight.

By Natalia Vargas, MPH

Most AI oversight asks whether a tool behaves safely. A second question gets far less attention. When computing capacity is tight, which AI workloads come first, and who decided?

Rural hospitals, health systems and public health agencies need a practical way to answer. The brief builds on the Themis Insights article, *Why Healthcare Leaders Must Govern AI Compute, Not Just Software*. The guidance is written for organizations that buy AI as a service and have no dedicated infrastructure team.

Why this matters now

Chips are scarce

A May 2026 report found that AI chip manufacturing is becoming a binding constraint on the pace of AI computing buildout [1]. A September 2026 analysis expects the AI-driven memory shortage to last through 2027 and possibly into 2028 [2].

Location matters

HIPAA permits storing protected health information outside the United States when a business associate agreement is in place. HHS cautions that the risks vary with geographic location [3].

Vendors should show their work

A 2026 sector guide urges health organizations to review where vendors host and process data, what uptime they commit to, and how they handle rollbacks [4].

A three-tier framework for sorting every AI workload

Sort each AI tool by the harm a delay would cause. Then write the rule that says who gives way first when capacity runs short.

Tier 1

Patient-adjacent and real-time

Delay can affect care or a clinician's time with a patient.

- Ambient documentation during visits
- Real-time clinical decision support
- Diagnostic imaging assist

Protect first. Write a priority rule, get performance commitments from the vendor, and test the fallback.

Tier 2

Operational and delay-tolerant

Valuable work that can wait minutes or hours.

- Prior authorization processing
- Patient message drafting
- Automated coding assist

Shared capacity is fine. The workload scales down automatically when Tier 1 needs room.

Tier 3

Administrative and batch

Back-office work with no patient or clinician waiting.

- Policy and report summaries
- Meeting transcripts
- Training material design

Run in off-peak windows. Pause this tier first.

General recommendations, not a prescription. The tiers above are a starting point. AI workloads differ by facility, so decisions should rest on a tailored framework built from an assessment of your own AI use cases. Clinical risk assessment needs precision, and your clinical safety committee, IT team and vendors should set the performance targets and run windows. **Request a tailored governance toolkit build →**

Three questions to ask every AI vendor

1 Location

Where will the work run? Name the country, the region and every subcontractor involved in hosting. Is a business associate agreement in place with each party that handles protected health information?

2 Resilience

What happens when a region fails or capacity runs short? Ask for failover and recovery timelines in writing, and for rollback support when an update causes problems.

3 Dependency and change

Which cloud and model providers sit underneath the service? How much notice will you receive before a model or service is retired? The sector guide suggests negotiating 12 to 18 months [4]. Could a smaller model meet the clinical need?

Your first step: the compute map

List every AI tool in use, including tools that ship inside your electronic health record. One page turns an invisible dependency into a managed risk.

Ready to go further? The Compute Governance Toolkit adds scoring, vendor questionnaires and contract language.
Request early access →

AI TOOL	WHO HOSTS IT	WHERE IT RUNS	PRIORITY TIER	FALLBACK IF CAPACITY FAILS	OWNER
Ambient documentation scribe	*Vendor, on a public cloud*	*Region named in contract*	*Tier 1*	*Manual note template*	*Chief medical information officer*

Example row shown in italics. Print this page or copy the table into your own document.

Beyond compute: four areas to score every AI vendor

Compute is one part of vendor risk. Score each vendor on four areas that map to the National Institute of Standards and Technology (NIST) AI Risk Management Framework [5] and the European Union (EU) AI Act [6]. Go deepest on Tier 1 tools. A short written attestation may be enough for Tier 3.

Data privacy and sovereignty

Are training datasets documented? What is the posture on customer data retention, and is a zero-data-retention option available? Does your organization track unsanctioned use of consumer AI tools, which can raise exposure of personally identifiable information?

Transparency and auditability

Does the vendor keep logs that show why the tool produced a specific output or took a specific action? Can your team retrieve those logs during an incident review?

Intellectual property and liability

Does the vendor offer indemnification for copyright claims tied to generated output? Can the vendor show where its foundation model came from and how it was trained?

Safety assurance

Ask for evidence of algorithmic bias monitoring, adversarial red-teaming results, baseline hallucination rates for your use case, and a formal vulnerability disclosure program.

Ready to put the framework to work?

The Compute Governance Toolkit turns this brief into working tools you can use this quarter:

- **Tier assignment worksheet** with a risk-scoring rubric and vendor review depth by tier
- **Vendor scorecard and questionnaire** scored on compute, data privacy, transparency, IP and safety assurance
- **Contract clause library** for location, recovery and notice of change
- **Surge and failover test checklist** to prove the fallback works
- **Board briefing one-pager** to brief executives and trustees
- **30-60-90 day rollout plan** with owners and checkpoints

Request early access, or ask about adding the toolkit to an Expert Advisory Hour.

Request the toolkit themisdigitalhealth.com/compute-governance-brief

About the author

Natalia Vargas, MPH, is the founder of Themis Digital Health. She served 13 years at the U.S. Department of Health and Human Services, including as a designee to the HHS AI Taskforce, and has completed more than 170 hours of targeted AI training in policy, legal considerations and AI-driven business strategy.

References

1. Sanders, J., Egan, J., & Madigan, R. (2026, May 7). *American AI companies can't get enough chips*. Center for a New American Security. https://www.cnas.org/publications/reports/american-ai-companies-cant-get-enough-chips
2. Shivakumar, S., Wessner, C., & Howell, T. (2026, September 17). *Beyond the memory cycle: AI, HBM, and the new semiconductor shortage*. Center for Strategic and International Studies. https://www.csis.org/analysis/beyond-memory-cycle-ai-hbm-and-new-semiconductor-shortage
3. U.S. Department of Health and Human Services, Office for Civil Rights. (2022). *Guidance on HIPAA and cloud computing*. https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html
4. Healthcare and Public Health Sector Coordinating Council, Cybersecurity Working Group. (2026, April). *Third-party AI risk and supply chain transparency guide*. https://healthsectorcouncil.org/wp-content/uploads/2026/04/AI-Third-Party-Risk-Guide.pdf
5. Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
6. Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Informational only. Not legal, clinical or compliance advice. The framework is a starting point for discussion with your clinical, legal and IT leaders.