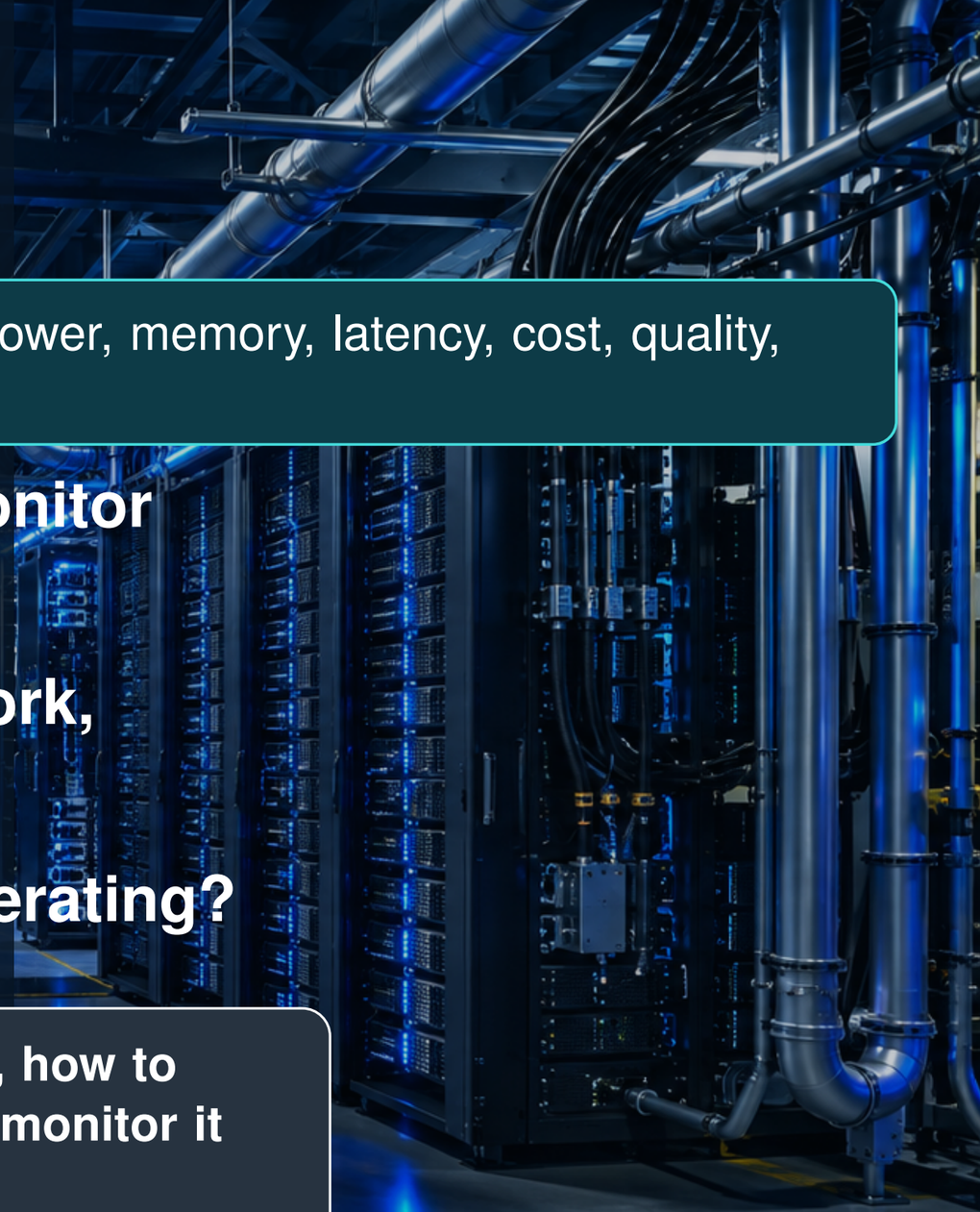
AI models are becoming data-center infrastructure



Takeaway: Model choice is now an infrastructure decision - power, memory, latency, cost, quality, and risk all matter.

- Enterprises will **run, fine-tune, distill, and monitor** their own models.
- AI workloads consume **GPUs, memory, network, power, and racks**.
- The real question: **Which model is worth operating?**

Four choices define the operating model: **which model to run, how to evaluate it, whether to use RAG or fine-tuning, and how to monitor it in production.**

Charles H. Martin, PhD

Founder, WeightWatcher | weightwatcher.ai

The AI workload spans tiny to trillion-scale models

Takeaway: There is no single best model tier - small, mid-size, large, and MoE models each play a different operational role.

Small
1B-8B

Local, private, fast, cheap.

Workhorse
30B-70B

Search, coding, support, knowledge work.

Frontier
200B+

Hard reasoning, teachers, evaluators.

Sparse MoE
1T+

Huge capacity, partial activation, complex serving.

Examples span Llama 3.1 405B, DeepSeek-V3 671B total / 37B active, and Kimi K2 1T total / 32B active. The right architecture is usually a **portfolio**, not one model.

Charles H. Martin, PhD

Founder, WeightWatcher | weightwatcher.ai

Model selection is now a supply-chain problem

Takeaway: There are too many plausible models to choose casually, so teams need a disciplined shortlisting process.

- The public ecosystem has **millions of model repositories**.
- Base models spawn **quantizations, adapters, distills, rerankers, embeddings, and forks**.
- Teams must filter by **task fit, hardware fit, cost, license, and proof path**.

Model selection should look less like picking a chatbot and more like **technical due diligence**.

Charles H. Martin, PhD

Founder, WeightWatcher | weightwatcher.ai

We are in the steamboat era of AI

Takeaway: Capability is scaling faster than our science of reliability. The systems work, but they can still fail in surprising ways.

- Early steam engines became useful before reliability science was mature.
- AI is similar: bigger systems often work better, but create **new instability and cost**.
- Failures include **hallucinations, weak retrieval, bad tool calls, and fine-tune regressions**.

The goal is not a demo that looks good once. The goal is a system **reliable enough, cheap enough, and observable enough** to run.

Charles H. Martin, PhD

Founder, WeightWatcher | weightwatcher.ai

Evaluation is the control system for enterprise AI

Takeaway: If you cannot evaluate a model for your actual task, you cannot manage it responsibly.

Public
Benchmarks

Broad signal, not production proof.

Custom
Task evals

Real workflows, outputs, edge cases.

RAG
Evals

Retrieval, grounding, evidence use.

Production
Evals

Drift, failures, cost, regressions.

Good evaluations are **scientific instruments**. They are harder to design than people expect - and often matter more than the model.

Charles H. Martin, PhD

Founder, WeightWatcher | weightwatcher.ai

RAG, fine-tune, distill, or train?

Takeaway: Use the simplest method that changes the thing that actually needs to change.

RAG

changes knowledge

Add current or private facts.

Fine-tune

changes behavior

Format, tone, tools, workflow.

Distill

changes cost

Smaller, faster, cheaper serving.

RL / Agents

change policy

Multi-step decisions and tool use.

Train

changes capability

New modality, IP, or sovereign control.

Decision rule: **RAG changes facts. Fine-tuning changes behavior. Distillation changes economics. Training changes capability.**

Charles H. Martin, PhD

Founder, WeightWatcher | weightwatcher.ai

The best architecture is usually a model portfolio

Takeaway: Do not force one giant model to do every job. Route work to the cheapest model that is good enough.

- **Small models** handle routine, private, low-latency work.
- **Large models** handle hard reasoning, supervision, and long-tail cases.
- **RAG, routing, and distillation** combine capability with efficiency.

A modern AI stack often uses **small models for routine work, large models for hard cases, RAG for facts, and distillation for scale**.

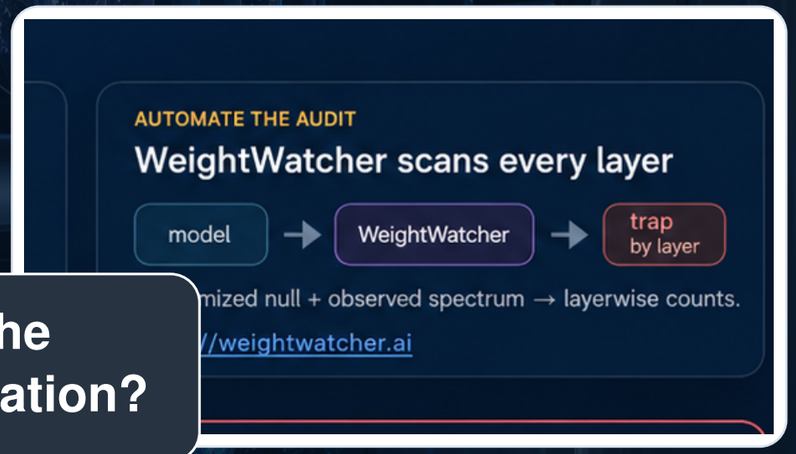
Charles H. Martin, PhD

Founder, WeightWatcher | weightwatcher.ai

WeightWatcher gives you an X-ray for trained models

Takeaway: WeightWatcher complements task evaluations with a fast model-health signal from the trained weights themselves.

- Inspect trained weights **layer by layer**.
- Compare **models, checkpoints, and tuning runs**.
- Flag signals tied to **overfitting, weak layers, instability, and structural quality**.



Task evaluation asks, **does it perform?** WeightWatcher asks, **does the model itself look healthy enough to trust and worth deeper evaluation?**

Charles H. Martin, PhD

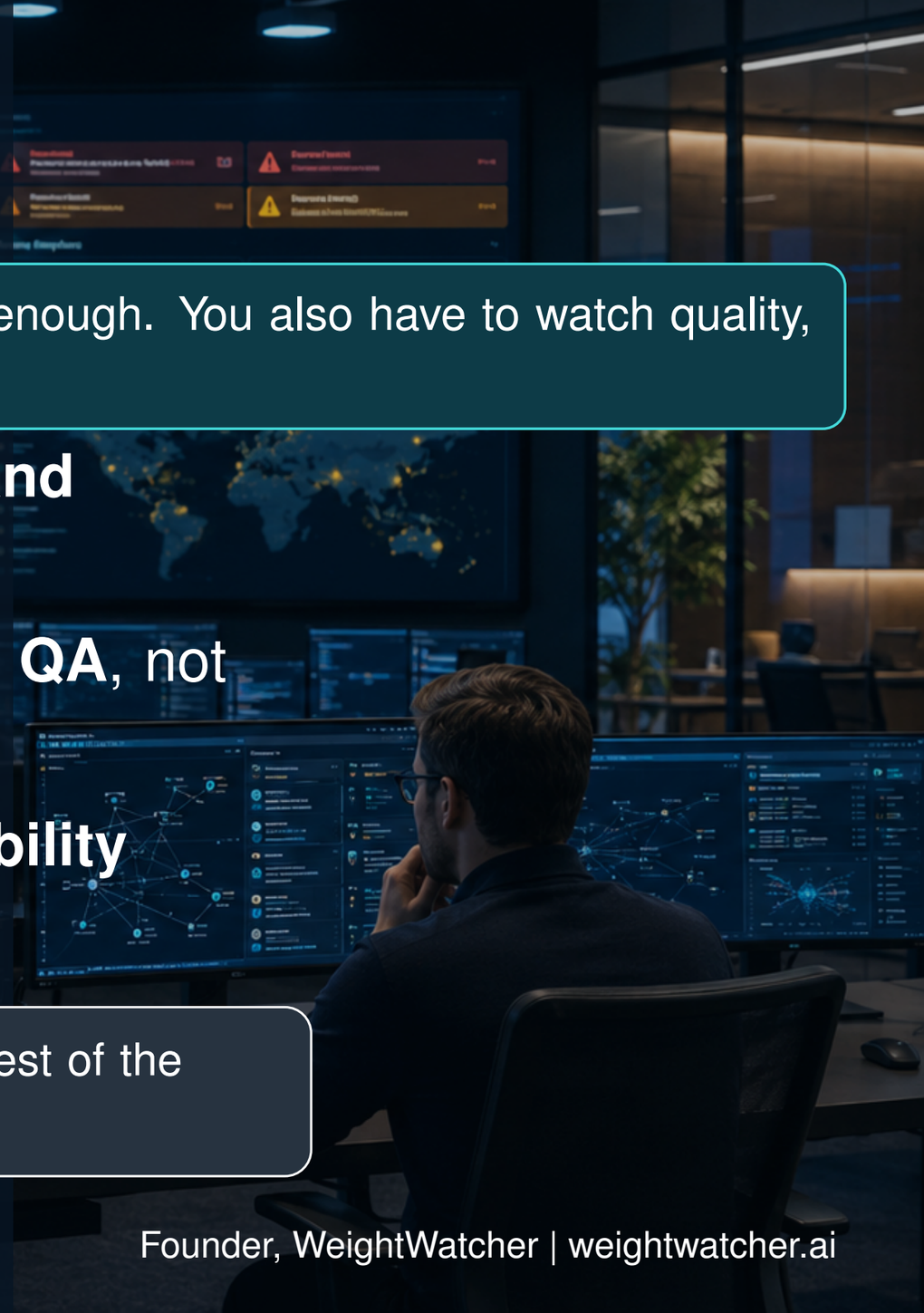
Founder, WeightWatcher | weightwatcher.ai

Production AI needs model-health monitoring

Takeaway: Once a model is live, uptime and latency are not enough. You also have to watch quality, drift, regressions, and cost.

- Monitor **quality drift, failure patterns, cost, and latency** over time.
- Fine-tuned and distilled models need **ongoing QA**, not one launch test.
- Diagnostics can catch **degradation and instability** before they become incidents.

AI infrastructure needs the same operational discipline as the rest of the data center: **measure it, tune it, and watch it continuously**.



Charles H. Martin, PhD

Founder, WeightWatcher | weightwatcher.ai

Enterprise AI needs a repeatable operating loop

Takeaway: The goal is useful work per watt, per dollar, and per unit of risk - not a one-time demo.

Choose

Shortlist by task, hardware, cost, privacy, and license.

Prove

Run task evals, RAG evals, red-team tests, and health checks.

Tune

Use RAG, fine-tuning, distillation, routing, or retraining.

Monitor

Track drift, regressions, cost, reliability, and model health.

If you are choosing, fine-tuning, distilling, or serving AI models, **I can help with model selection, evaluation design, WeightWatcher audits, and AI roadmap decisions.**

Charles H. Martin, PhD

Founder, WeightWatcher | weightwatcher.ai