

THE HARDWARE MOVED

A trillion-parameter model now fits on a desk.

That changes how you adopt agentic AI, and the three decisions that decide whether it works have nothing to do with the hardware.

Ten slides, five minutes.

- 01** Agents arrive at the desk and at the edge
 - 02** A robot is an agent with a physical presence
 - 03** The frameworks that made this practical
 - 04** Where a workload runs is now a cost decision
 - 05** People, economics, operating model
-

748GB

of coherent memory in a deskside machine, enough to run frontier models of up to one trillion parameters locally.

NVIDIA DGX Station for Windows, announced May 2026, shipping through OEM partners from Q4 2026.

\$3,999

launch price of DGX Spark, with 128GB of unified memory and inference on models up to 200 billion parameters.

Two to four units cluster to 256GB and 512GB. In twelve months the deskside ceiling went from tens of billions of parameters to a trillion.

A robot is an agent with a physical presence.

- **Same loop.**

Perceive, plan, act. The difference is that the action moves an object in the world rather than a row in a database.

- **Same policy.**

You do not need a robotics policy and an agent policy. You need one set of autonomy tiers.

- **Different blast radius.**

A picking robot that can injure someone belongs at a lower tier than a scheduling agent that can annoy someone.

Long-running agents became a download.

FRAMEWORK	WHAT IT BETS ON
OpenClaw	Runs on your own device across the messaging channels people already use. Hosted and local models are interchangeable. Past 389,000 GitHub stars.
Hermes Agent	A learning loop. It writes reusable skill documents from experience and approaches the same task differently next time. Nous Research, February 2026.
NemoClaw	The security layer. Sandboxed execution, local Nemotron models, a privacy router and default-deny networking. NVIDIA, March 2026.

Three placements, deliberately assigned.

PLACEMENT	WINS	FAILS
Cloud API	Variable volume, frontier capability, no infra team	Always-on agents, data residency, control loops
Local or deskside	Continuous volume, sensitive data, low latency	Bursty work, no one to maintain it
Edge device	Physical sites, poor connectivity, milliseconds	Very large models, frequent updates across sites

The mistake is choosing one on principle. Most organisations that get this right end up running all three.

//

The hardware is the easy part now. The person who knows why a process is shaped the way it is remains the scarce input.

TERENCE KOK

What actually decides adoption.

— People.

Name three super users. Deep process knowledge plus genuine openness to change. Those two traits rarely sit together, and the people who have both are your fastest route to production.

— Economics.

Cost the workload at the volume it reaches if the pilot works, not at pilot volume. Most cost surprises are volume rises nobody modelled.

— Operating model.

Adding agents to the structure you already have gives you a real but small gain, available to every competitor. Redesigning the structure is where the step change lives.

The technology is settled enough to plan around.

Whether your organisation has the people, the cost discipline and the appetite for structural change is the open question. None of those three are downloadable.



Terence Kok

AI & Innovation Strategist

SAVE · COMMENT · FOLLOW

Full article and free tools at terencekok.com