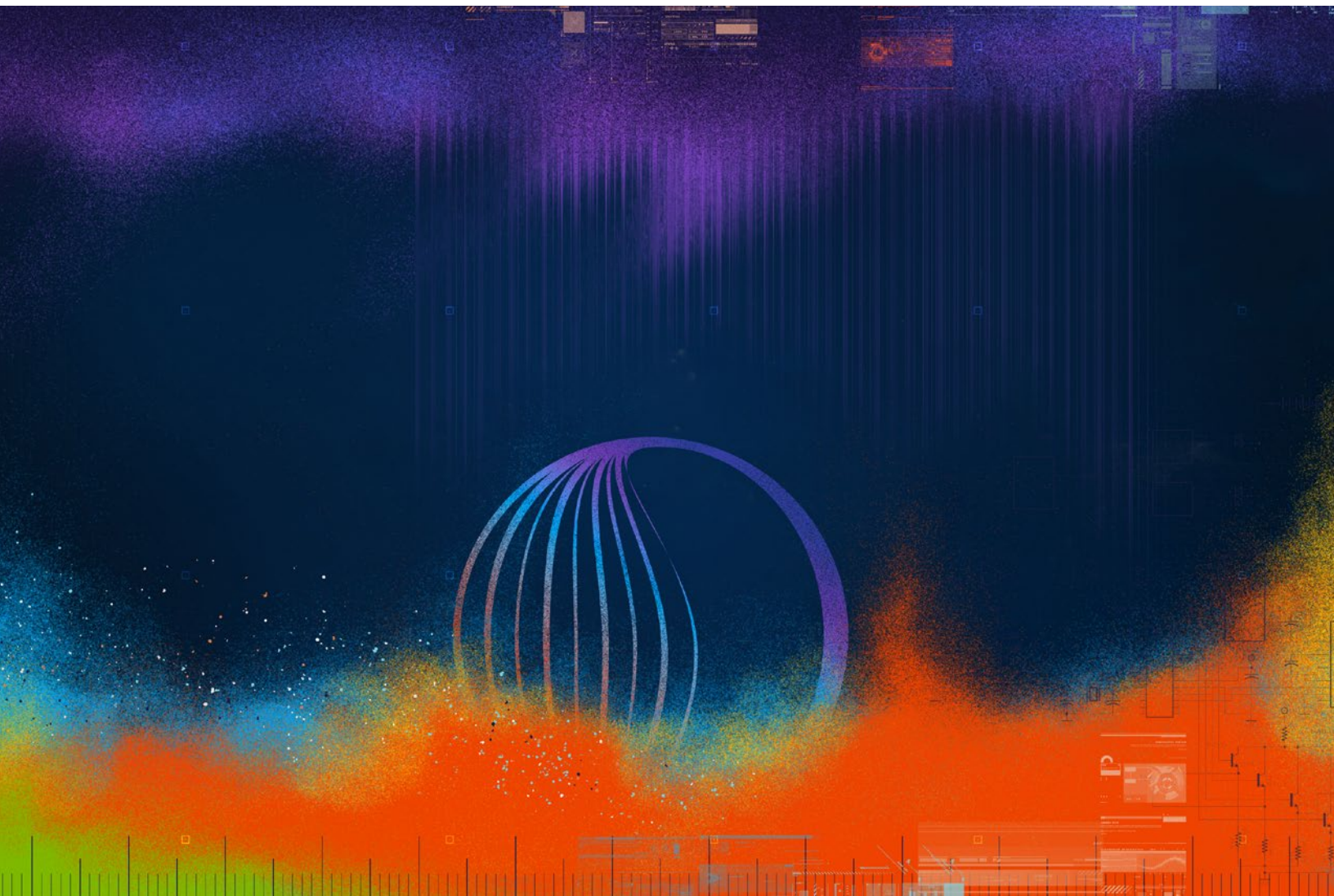




Unmetered Intelligence

How Windows Is Building the
Operating System for the Agentic Era,
Designed Around How People
and Enterprises Use AI

AUTHOR

Ryan Shrout

President & GM | Signal65

COMMISSIONED BY



SEPTEMBER 2026

Introduction

A forward-looking view of what an operating system must deliver as agentic workflows take hold, and where Windows stands today, grounded in what Microsoft announced at Build 2026.

The operating system is being rebuilt for how people and enterprises will work in the agentic era, and Windows is leading that shift with capability that is shipping or rolling out today. The throughline is unmetered intelligence, running AI locally and escalating to the cloud when the work calls for it, which changes the cost, privacy, and responsiveness of the AI experiences people rely on. This report defines the capabilities an operating system will need to deliver as agentic workflows take hold, shows where Windows stands against each one based on what Microsoft announced at Build 2026, and describes how that value scales across the device portfolio, from mainstream Copilot+ PCs to a deskside AI supercomputer. Throughout, we use the term local models for models that run on the device itself, a category also described as on-device models.

Key Highlights



Enterprise model API spend more than doubled to about \$8.4 billion by mid-2025, even as token prices fell



Local models on Windows, including Aion 1.0 Instruct and Plan, run agentic reasoning with no per-token cost



Routing work by complexity across local and cloud can cut inference cost by roughly 5 to 10 times



OS-enforced identity and containment let enterprises observe, govern, and secure every agent

- **The shift is human-first.** The design goal is not AI for its own sake. It is to give people and organizations optimized AI experiences and to make them ready for the agentic era, with the operating system evolving to serve how humans work.
- **Unmetered intelligence is the value.** Running the right work locally, without per-token cost or a cloud round trip is where newer, more capable hardware paired with Windows creates value a cloud-only model cannot match, especially as always-on agents make continuous inference the norm.
- **Governance is the unlock.** Windows lives up to the enterprise promise by letting organizations observe, govern, and secure agents at the operating-system layer, so agent adoption can scale with control and trust intact.
- **It scales across every tier.** The same Windows AI platform runs from the broad Copilot+ PC install base to high-performance and frontier systems, so the value reaches mainstream users and frontier developers alike.

This report was commissioned by Microsoft and authored by Signal65. The analysis, findings, and conclusions expressed in this report are those of Signal65. Microsoft provided publicly available information regarding Build 2026 announcements and reviewed the report for factual product accuracy and terminology but did not influence the report's analysis, findings, or conclusions.



The Signal65 Agentic OS Value Framework

Signal65 defines six capabilities that determine where an agentic, AI-centric operating system creates value, and Windows is delivering against all six today. The framework is our current view as of mid-2026 and is built to adapt as the category matures. It is a general model for evaluating any operating system built for agentic work, and it organizes the three pillars of this report into a single picture so a reader sees the whole opportunity at a glance.

Capability	Availability Status on Windows	Pillar
Unmetered Local Intelligence Local AI that runs continuously without per-token cost or a cloud round trip.	Shipping / coming. Aion 1.0 Instruct and Aion 1.0 Plan local SLMs and coding models such as MAI-Code-1-Flash, plus Windows AI APIs expanding across CPU, GPU, and NPU, with Windows ML as the platform for local AI.	1
Hybrid Execution Intelligence placed where it makes the most sense, local and cloud together, by task.	Shipping. Delegation and routing of work from the cloud to local models by complexity, one implementation being the fleet capability shown at Build 2026, cutting cost without sacrificing quality.	1 2
Agentic Execution and Governance Agents that act, contained and attributable, with OS-enforced identity and policy.	In preview. Microsoft Execution Containers provide a dynamically composable containment spectrum; agents run with their own OS-enforced identity and Agent 365 governance through Defender, Entra, Intune, and Purview.	2
Open Agent Ecosystem The trusted place to run the agents and runtimes developers adopt.	Shipping / preview. OpenClaw runs contained on Windows via MXC, GitHub Copilot CLI has adopted MXC process isolation, and NVIDIA OpenShell builds on the same foundation; partners include Hermes, Manus, NVIDIA, OpenAI, and OpenClaw.	2
Platform Breadth Across Tiers One platform and model that scales from mainstream laptops to a deskside supercomputer.	Shipping / coming. Copilot+ PC install base, the Surface RTX Spark Dev Box and AMD Ryzen AI Max systems, and DGX Station for Windows, all running the same Windows AI platform.	3
Security and Identity by Default A platform that reduces risk by default so agents inherit protection without extra work.	Established. Hardware root of trust and passwordless sign-in, with agents observed via identity, governed via Agent 365, and secured via MXC, plus enterprise manageability through Intune and Entra.	2 3

Table 1: The Signal65 Agentic OS Value Framework. Six capabilities, what Windows delivers today against each, and the report pillar each maps to. Every entry is anchored to a Build 2026 announcement.

The Operating System Is Being Rebuilt Around How People Use AI, Now

For most of computing history, the operating system has been a launcher for applications. That role is changing. As AI moves from a feature inside individual apps to a capability woven through the whole system, the operating system is becoming the place where intelligence runs and where agents do work on a person's behalf. This is underway in 2026, grounded in capability that is shipping, and the design intent is squarely human-first, with the platform evolving so that people and enterprises can get more done with AI rather than so that AI can run for its own sake.

The era is here, not around the corner. A concrete signal is that developers can run **OpenClaw**, a widely adopted open-source agent runtime, securely on Windows today, with its node and gateway running inside operating-system-enforced containment. Windows is already hosting the agents developers are adopting, which is a very different thing from describing them for a later release.

Unmetered intelligence is the throughline of this report. As models grow more capable and agentic workflows demand continuous compute, moving the right work to the device changes the economics, privacy, and responsiveness of AI, so frontier models can tackle frontier problems while a large and growing share of everyday intelligence runs locally at scale and at no per-token cost. The sections that follow organize the case around three pillars, the economics of unmetered local intelligence, the way agentic workloads reshape the platform and why governance is the unlock, and how the value scales across the full device portfolio.

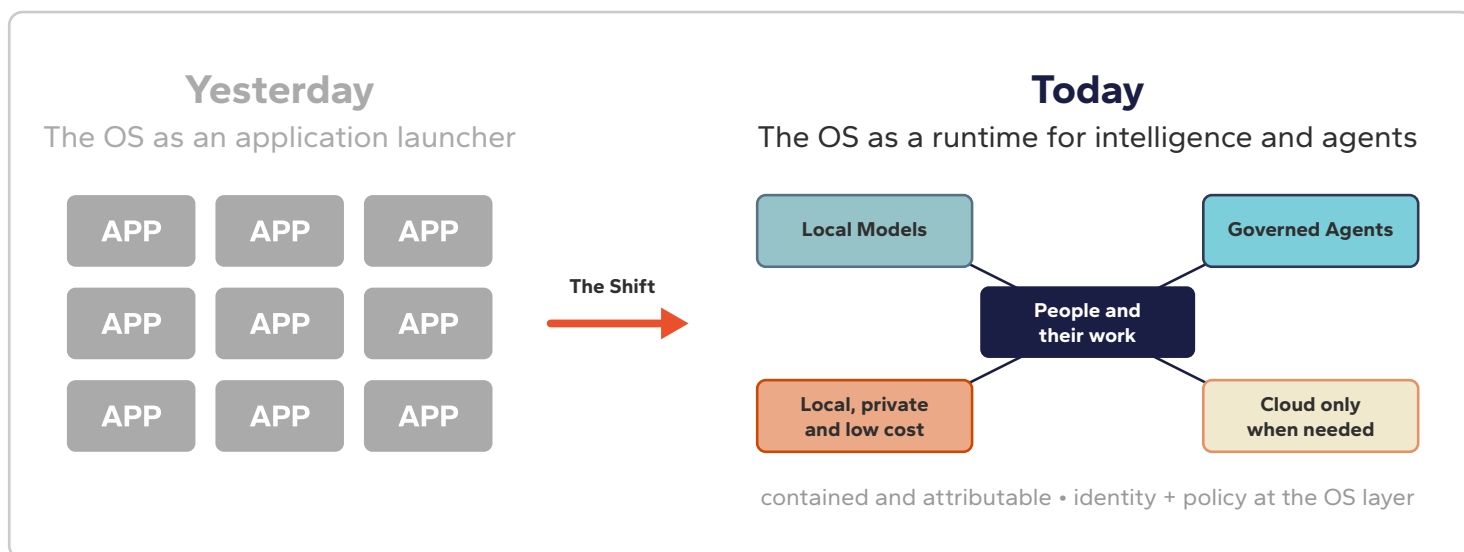


Figure 1: The shift from an operating system that launches applications to one that runs intelligence and governs agents, with people and their work at the center.

This report was commissioned by Microsoft and authored by Signal65. The analysis, findings, and conclusions expressed in this report are those of Signal65. Microsoft provided publicly available information regarding Build 2026 announcements and reviewed the report for factual product accuracy and terminology but did not influence the report's analysis, findings, or conclusions.

Unmetered Intelligence and the Economics of Local AI

1

Pillar 1. The cost and value case for running AI locally rather than cloud-only.

The Cost of Cloud-only Intelligence

Even as the price of a single token falls sharply, total spending on AI is rising, because usage is growing faster than prices are dropping. Enterprise spending on model APIs more than doubled in a single stretch, from roughly \$3.5 billion in late 2024 to about \$8.4 billion by mid-2025, according to [Menlo Ventures](#), which also found that nearly half of large enterprises now say most or nearly all of their compute is inference-driven, up from under a third a year earlier. Cost-tracking data from [CloudZero](#) points the same way, with average monthly AI spend climbing from about \$63,000 to about \$85,500 over a year and the share of companies planning to spend more than \$100,000 a month more than doubling. Agentic workflows are the main driver, because an agent reasons in a loop, calls tools, and chains many model calls to finish a single request, and always-on agents run continuously, which turns inference into a standing cost rather than a one-time one.

How Windows Changes the Equation

Windows answers the cost problem by making local execution a first-class path. At Build 2026 Microsoft introduced a new generation of local small language models for Windows, [Aion 1.0 Instruct](#), a smaller and faster local model, and Aion 1.0 Plan, a reasoning and tool-calling model that enables fully local agentic capabilities, both slated to arrive in the coming months. Paired with coding models such as [MAI-Code-1-Flash](#) that run on client devices, these let everyday text intelligence and local agentic reasoning run in-box on capable devices with no cloud dependency and no per-token cost, which is where unmetered value is realized rather than through cloud APIs alone. Microsoft is also expanding the [Windows AI APIs](#) beyond the NPU to run across CPU and GPU, with speech-to-text on NPUs and CPUs, the local model extending to capable discrete GPUs, and video super resolution on CPUs, so developers can deliver richer experiences without a cloud round trip. Windows ML provides the common runtime that lets a single model target the right silicon across the fleet.

The most practical way to control cost is to place each unit of work where it belongs. Windows supports selective delegation and routing of work between the cloud and the device. One implementation is the fleet capability Microsoft showed at Build 2026, where a cloud planner delegates suitable work down to local models by complexity, and other approaches are possible on the same platform. Cost analysis from [CloudZero](#) and related industry research finds that complexity-based routing and delegation, combined with caching and batching, can reduce total inference cost by roughly 5 to 10 times versus sending everything to a frontier cloud model. Framed for buyers, this is cost avoided, not a hardware specification.

A Second Dividend: Your Data, and the Learning it Creates, Stay Yours

Cost is not the only reason to run intelligence locally. There is a second and more strategic reason, which is keeping proprietary data, and the learning it generates, inside the organization's control. The economist Kenneth Arrow described a classic information paradox, in which a seller of information risks giving it away in the act of selling it. AI inverts the problem for the buyer. To get a model to perform well on your work, you have to feed it the proprietary context that makes your organization distinct, so you effectively pay for intelligence twice, once in money and again in the knowledge you must reveal to make it useful.

That exposure compounds. Models improve from the exhaust of real use, the prompts people write, the tools agents call, and especially the corrections people make when a model is wrong, and each of those is distilled institutional know-how that a competitor could not simply buy. Sent to a cloud provider, that know-how can leak imperceptibly, trace by trace and correction by correction. Running intelligence locally on Windows, inside the enterprise trust boundary, lets that data, those traces and evals, and any adapted model weights accumulate and compound where the organization controls them, rather than flowing to the owner of the cloud infrastructure. In the cloud era enterprises accumulated data. In the agentic era they accumulate learning, and the trust boundary has to evolve to protect the mechanism by which an organization learns and improves. Unmetered local intelligence and hybrid execution on Windows are the substrate for keeping that learning loop inside the organization.

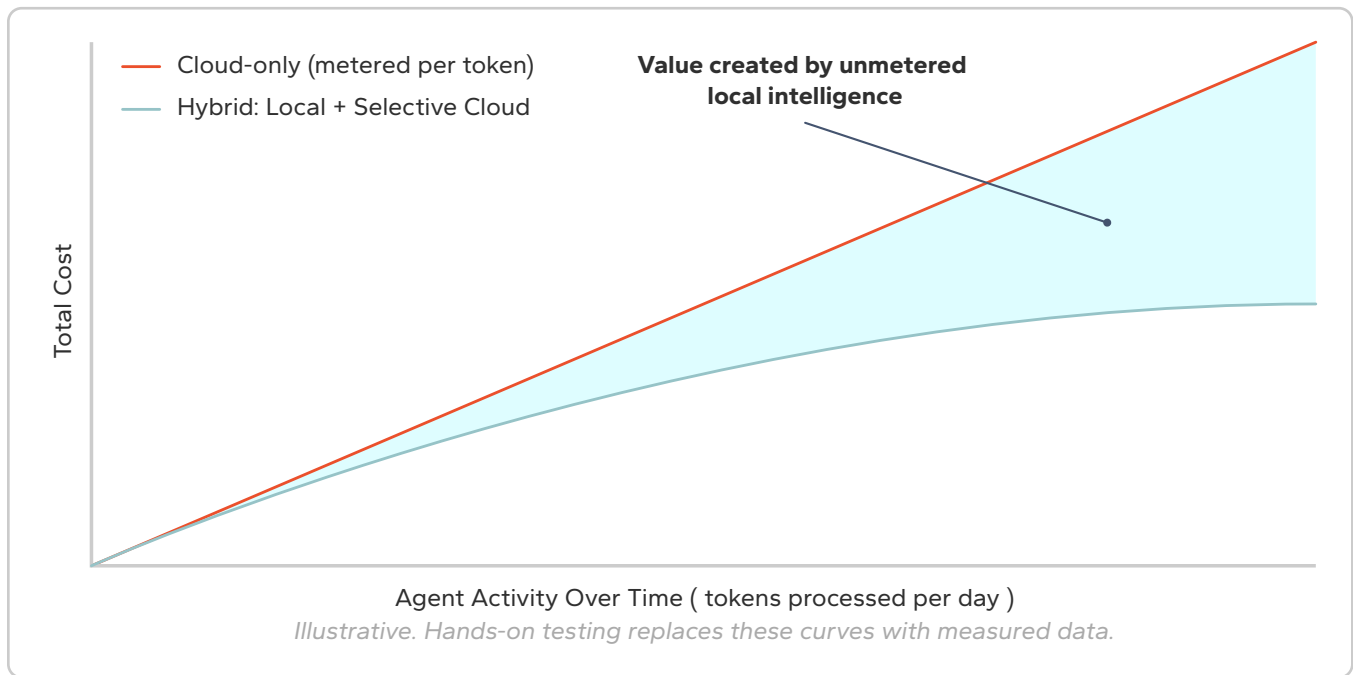


Figure 2: The cost of intelligence. A cloud-only, metered approach rises with agent activity, while a hybrid approach that runs the right work locally flattens the curve. The widening gap is the value created by unmetered local intelligence. Illustrative; hands-on testing in a follow-on phase would replace these curves with measured data.

Quantifying the Value

The value shows up as tokens and cost avoided by running routine, repetitive, and sustained work locally, whether that is the same task run many times with different parameters or many different tasks running over a long period, plus the privacy, data-control, and latency benefits of keeping work on the device. This paper presents the economic logic and publicly reported figures, and Signal65's upcoming hardware testing turns it into measured, defensible numbers on representative Windows systems.

Agentic Computing and the Evolution of the Platform



Pillar 2. How agentic workloads reshape the platform, and why governance is the unlock.

From Assistant to Actor

Agents no longer just answer questions. They read and write files, invoke services, modify environments, and chain operations together at speed. That capability is powerful, and it raises a trust question the operating system has to answer, which is how these systems stay trustworthy when they act autonomously, at scale, on real data. This is precisely where Windows differentiates, by building the answer into the platform rather than leaving it to each application.

The Enterprise Promise: Observe, Govern, Secure

The platform gives organizations three foundations for running agents with confidence. It lets them observe, through identity, so every agent's activity is attributable and human and agent can be told apart. It lets them govern, through **Agent 365** integration, so policy and lifecycle management apply to agents the way they do to users. And it lets them secure, through **Microsoft Execution Containers**, so what an agent can touch is bounded at runtime. Together these let agent adoption scale while control and trust stay intact.

Containment and Identity, Enforced by the OS

Microsoft Execution Containers, in early preview, provide a policy-driven execution layer where developers declare what an agent may access and Windows enforces those limits at runtime. The isolation is a dynamically composable containment spectrum, from lightweight process isolation for a responsive coding inner loop, to session isolation that separates an agent's desktop, clipboard, and input from the user's, to cloud-isolated execution through **Windows 365 for Agents**, all through one policy model and SDK, so a coding agent and an enterprise data-processing agent can have different guardrails but one coherent trust story.

Underneath, Windows assigns each agent a local identity, or a cloud-provisioned identity backed by Microsoft Entra, and attributes all of its activity to that identity, which enables least-privilege access and auditability. Agent 365 then applies governance and manageability through Defender, Entra, Intune, and Purview, so agents start secure and stay secure and security teams can constrain a local agent before it becomes enterprise risk, with Defender extending protection to emerging agent threats such as prompt injection.

The Trusted Place to Run the Ecosystem

The same foundation runs the agents developers actually adopt. OpenClaw runs its node and gateway contained on Windows through MXC, GitHub Copilot CLI has adopted MXC process isolation to constrain what its generated code can do, and NVIDIA is bringing OpenShell to Windows built on the same foundation for always-on agents. Microsoft is partnering with a broad set of ecosystem leaders on this containment model, including Hermes, Manus, NVIDIA, OpenAI, and OpenClaw. Local agents handle private, low-cost loops on the device, while heavier reasoning and higher-risk work escalate to the cloud, including Azure, or to an Intune-managed Cloud PC through Windows 365 for Agents, the local-executor-with-cloud-escalation pattern the platform is built around, which ties the governance story back to the unmetered-intelligence economics of Pillar 1.

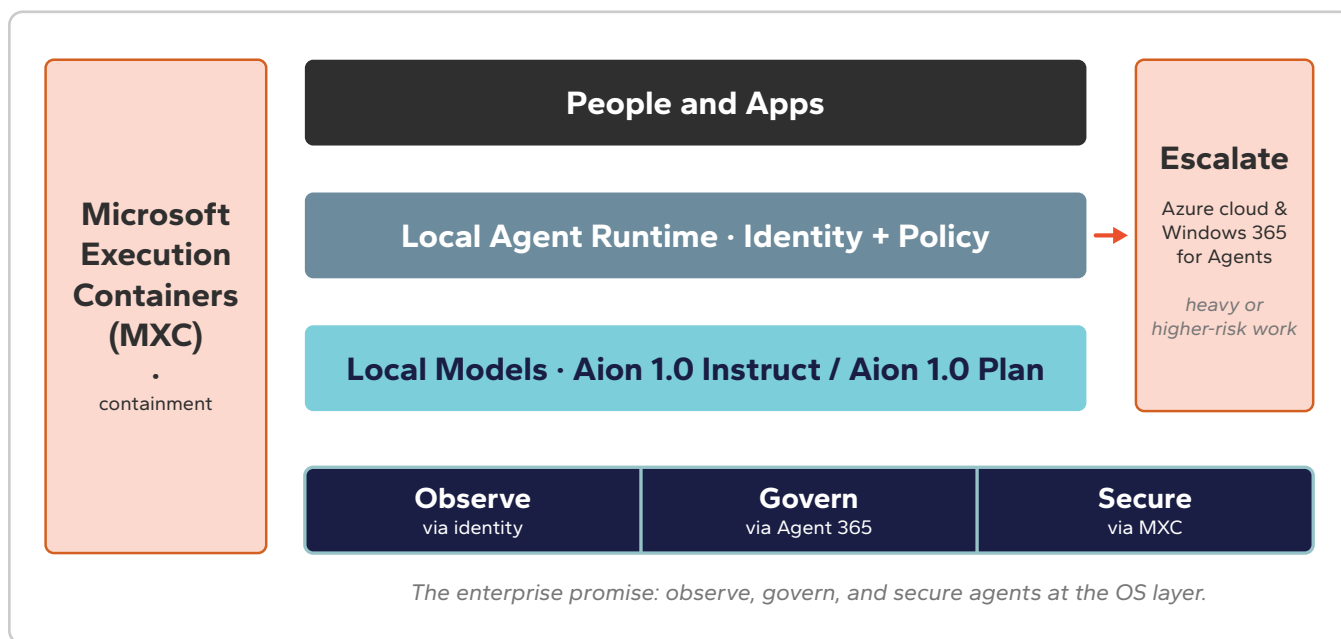


Figure 3: The Windows agent platform: people and apps on top, a local agent runtime carrying identity and policy, local models beneath it, and escalation to Azure cloud and Windows 365 for Agents, all wrapped by Microsoft Execution Containers, with the enterprise foundations of observe, govern, and secure.

One Platform, Every Tier: Windows Breadth



Pillar 3. How the value scales across the device portfolio.

The value of unmetered local intelligence and governed agents is not confined to one class of machine. Because the same Windows AI platform, the same containment and identity model, and the same developer experience run across the portfolio, the value scales smoothly from a mainstream laptop to a desktide supercomputer. That breadth is a structural strength of Windows, and it spans three tiers.

Tier	Representative Windows Systems	What it Unlocks
Mainstream	The broad Copilot+ PC install base	Puts local AI into everyday hands. The Aion models and the expanded Windows AI APIs bring unmetered local intelligence to a large and growing base of capable machines from many manufacturers.
High-performance	Surface RTX Spark Dev Box (NVIDIA RTX Spark, up to 1 petaflop, 128GB unified memory), and AMD Ryzen AI Max systems (Ryzen AI Max+ 395 and the Ryzen AI Halo developer platform, up to 128GB unified memory)	Makes local model optimization, fine-tuning, and large local inference practical, running 100B-plus-parameter models on a single system on Windows, without recurring cloud cost.
Frontier	DGX Station for Windows (NVIDIA GB300 Grace Blackwell Ultra, frontier models up to 1 trillion parameters), and workstations configured with multiple NVIDIA RTX PRO 6000 Blackwell GPUs (two to four cards pooling 192 to 384GB)	Develops and runs frontier-scale models locally and connects always-on, frontier agents to enterprise workflows, at the top of the portfolio.

Table 2: Windows across three device tiers, with representative systems and what each unlocks for local AI. Device and pricing detail is kept intentionally high level in this report.

Consistency is what turns a set of devices into a platform. A developer can build and govern an agent the same way on a mainstream Copilot+ PC, a high-performance box like the upcoming Surface RTX Spark Dev Box or an AMD Ryzen AI Max system, and a frontier machine like DGX Station for Windows, and an organization can apply the same identity and policy model across all three. The portfolio spans a wide range of price points, from roughly \$800 to \$2,500 at the mainstream entry level to roughly \$3,500 to \$50,000 and up at the high-performance and frontier tiers, so organizations can match the level of local capability to the work in front of them.

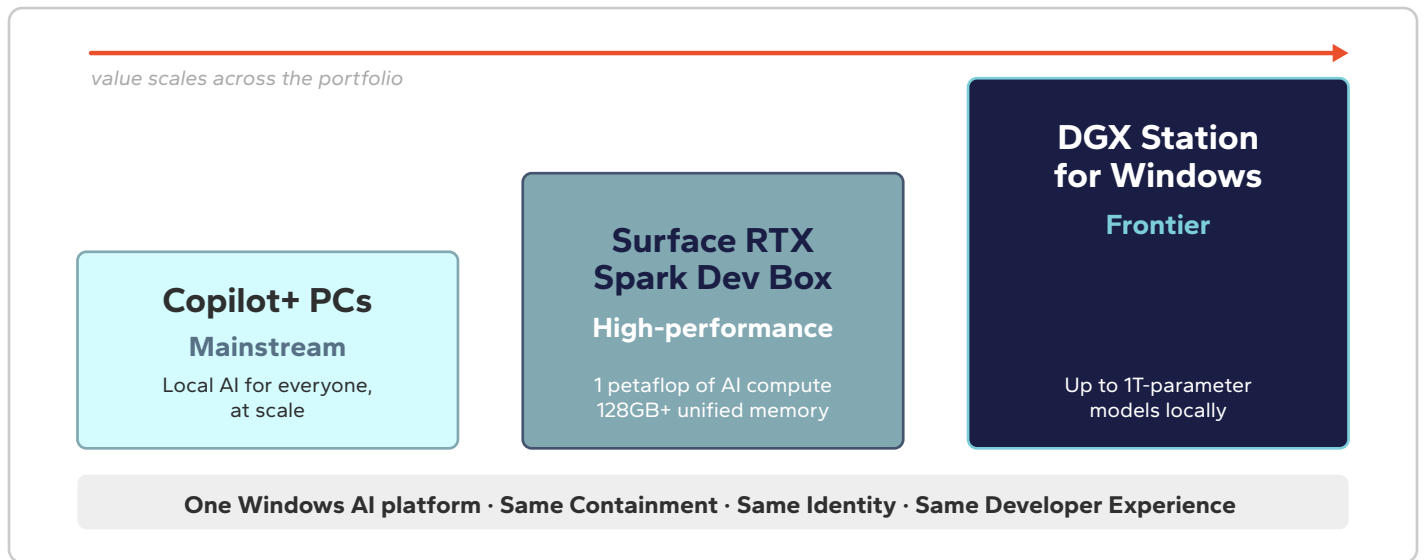


Figure 4: One platform, every tier. The same Windows AI platform, containment, and identity run from mainstream Copilot+ PCs through high-performance and frontier systems, so the value scales across the portfolio.

Where Windows Is Headed, and What Comes Next

Windows is building the operating system for how people and enterprises will use AI, pairing unmetered local intelligence and governed agents with a portfolio that scales across every tier, on a foundation that already runs the world's workflows. The pieces announced at Build 2026 fit together into one story, running the right intelligence in the right place, giving organizations the ability to observe, govern, and secure the agents doing the work, and making that value available from the most mainstream device to the most powerful.

The most reliable way through a period of rapid change is to layer new intelligence onto what already works, rather than asking people and organizations to relearn how they use their devices. Windows is taking that additive path, extending a familiar, broadly compatible platform into the agentic era rather than replacing it.

As next-generation hardware reaches the market this year, Signal65 will complement this analysis with hands-on testing that quantifies the unmetered-intelligence and hybrid-execution value with measured results across the device tiers, so the cost-avoided value rests on data, not only economic logic.

The agentic era is still taking shape, and the framework in this report is built to adapt as it does. Our aim is to describe where the platform is heading with confidence where capability is real and announced, and with appropriate humility about what is still to come.



Selected Sources

Microsoft, [Build 2026: Furthering Windows as the trusted platform for development](#) (Windows Developer Blog, June 2, 2026).

Microsoft, [Windows platform security for AI agents](#) (Windows Developer Blog, June 2, 2026).

Microsoft AI, [Introducing MAI-Code-1-Flash](#); Microsoft Execution Containers (MXC) SDK, github.com/microsoft/mxc.

NVIDIA, [RTX PRO 6000 Blackwell Workstation Edition](#); [NVIDIA DGX Station](#).

AMD, [Ryzen AI Max and the Ryzen AI Halo developer platform](#) (AMD, 2026).

Menlo Ventures, [2025 Mid-Year LLM Market Update](#) (enterprise model-API spend \$3.5B to \$8.4B; inference share of enterprise compute).

CloudZero, [State of AI Costs](#) (average monthly AI spend \$63K in 2024 to \$85.5K in 2025; cost analysis underpinning the routing and delegation savings range).

Important Information About this Report

AUTHOR

Ryan Shrout

President, GM | Signal65

PUBLISHER

Ryan Shrout

President, GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.



CONTACT INFORMATION

Signal65 | signal65.com