

Home • Artificial Intelligence

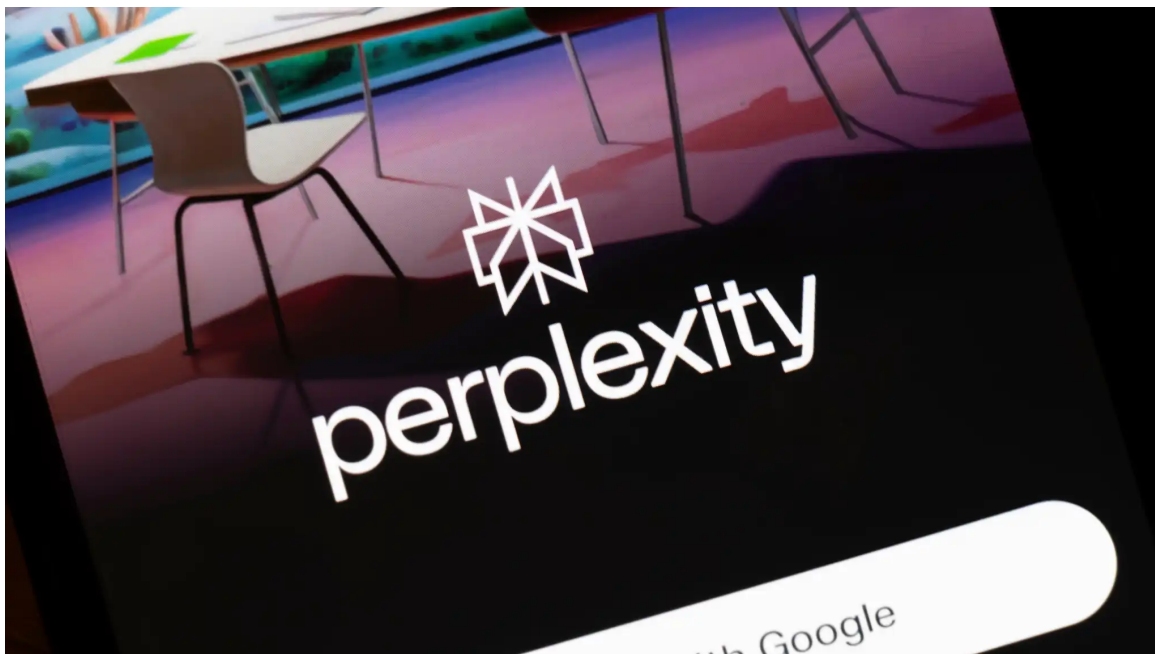
by **Evan Schuman**
Contributor

Perplexity's on-device AI offering promises data control and lower token costs

News

Aug 25, 2026 • 7 mins

But some question how well that data control is being enforced; hardware cost is also an issue



Credit: daily_creativity / shutterstock

Perplexity on Tuesday rolled out an offering that runs the AI entirely on a local machine, and that, it said, will keep “private data local and escalating to the cloud only when a task needs it.”

The service, called simply Portable Computer, is a local version of **Perplexity Computer** [<https://www.computerworld.com/article/4138723/perplexitys-new-computer-agent-will-run-other-agents-for-you.html>] that runs on the Nvidia DGX Spark with Qwen 3.8 27B or with PPLX 27B, a post-trained version of the Qwen model, Perplexity said, adding that a 30B open model is coming soon to the model picker. The orchestrator, planner, tool router, scheduler, durable task queue, and local search index all run on device.

Portable Computer requires Linux as its underlying OS, with Windows support “coming soon.”

The most critical benefits for enterprise IT are the Perplexity promise that data stays local, and that on-device work doesn’t consume token credits. Customers are only charged if the system is explicitly told to move compute to the cloud “for more advanced research and reasoning.”

Doing as much work as possible locally is **becoming a trend** [<https://www.computerworld.com/article/4095833/microsofts-fara-7b-brings-ai-agents-to-the-pc-with-on-device-automation.html>], as users seek to lower token costs and minimize their need for **increasingly expensive high-end AI systems** [<https://www.cio.com/article/4213246/nvidia-to-hike-prices-by-15-on-top-of-an-even-larger-increase-in-july-2.html>].

“A user might want the confidential details of a term sheet to stay local, but escalate to the cloud to get current market comps or recent precedent deals,” the **Perplexity announcement** [<https://www.perplexity.ai/hub/blog/introducing-portable-computer-for-local-first-ai>] noted. “Portable Computer’s local orchestrator can escalate a task to the cloud for current information, browser use, connected apps, or one of 15+ frontier models for advanced reasoning. Through app connectors, Portable Computer works with Google Drive, Gmail, Slack, and GitHub.”

Recommended for you



Microsoft is turning Windows 11 into an AI operating system
By Agam Shah • 4 mins

[https://www.computerworld.com/?post_type=post&p=4189045&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]



IT infrastructure shortages are real

[https://www.cio.com/?post_type=post&p=4215445&utm_source=miso&utm_medium=related]

and lasting. Here's
how to cope

By Paul Desmond • 13 mins



Macs don't just do
AI, they're
replacing the cloud
for it

By Jonny Evans • 5 mins

d&utm_campaign=thumbnail_list]

[https://www.computerworld.com/?post_type=post&p=4218761&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]

Aman Mahapatra [<https://www.linkedin.com/in/akm76/>], chief strategy officer for Tribeca Softtech, a New York City-based technology consulting firm, said that he liked the comprehensive capabilities included in the offering.

“Running a model locally has been table stakes for two years,” Mahapatra said. “Running the entire agentic control plane locally means the decision about whether a task needs the cloud is itself made on device, by a model post-trained to keep work local and escalate when necessary.”

Hardware demands

However, some consultants and analysts questioned whether the economics, as well as the data and token control features, are truly ready for the enterprise at this point.

Nader Henein [<https://www.gartner.com/en/experts/nader-henein>], a Gartner VP analyst, said that some of the announced Perplexity models can be run today on high-end laptops with off-the-shelf GPUs, “but until we see the price [of Portable Computer], it’s going to be hard to get excited about this.”

Computerworld Smart Answers [Learn more](#)

Explore related questions

- [What are the essential steps for implementing private cloud AI?](#)
- [Why is sandboxing local AI agents an operational challenge for enterprises?](#)
- [How does TransferEngine help enterprises avoid AI vendor lock-in?](#)
- [How does quantization reduce memory requirements for local AI models?](#)
- [Why are small language models reducing enterprise reliance on cloud AI?](#)

Flavio Villanustre [<https://www.linkedin.com/in/fvillanustre/>], CISO for the LexisNexis Risk Solutions Group, also noted, “the hardware demands are quite steep, especially with the current hardware costs including RAM, GPU, and so forth. It needs at least a local GPU with 24GB of VRAM as a minimum.”

“Although it may help lower ongoing expenses, it does require an initial investment on this specialized hardware,” he said.

Concerns over control of cloud usage

But multiple consultants expressed concerns about the lack of details of how the local-versus-cloud decisions are enforced. For example, users might agree to a pop-up offering cloud compute reflexively, just as they simply agree to any software terms and conditions that pop up. Or an attacker could use prompt engineering or other tactics to place hidden commands that tell the system to move to the cloud without the user knowing.

Justin Greis [<https://acceligence.com/talent/profiles/justin-greis/>], CEO of consulting firm Acceligence, pointed out, “local-first should not be confused with local-only and that distinction is going to matter tremendously. Users routinely approve prompts they do not fully understand, and increasingly autonomous AI agents are operating across files, applications, connectors, and workflows that may be far more complicated than the user can see. The AI should be able to ask for permission, but the enterprise needs the ability to say, ‘You are not permitted to ask.’”

Mike Wilkes [<https://www.linkedin.com/in/eclectiquis/>], enterprise CISO at Aikido Security, suggested that one way to prevent cloud usage would be for IT to lock down these systems so that they cannot have any external access, a technique that he doubted would be tolerated.

“I would not expect either the agents or their users to be particularly happy if enterprises solve the security problem simply by denying network access. For many valuable use cases, connectivity is the point,” Wilkes said. “A proprietary trading firm, for example, may want confidential models and positions processed locally while still consuming current market data, SEC filings, news or research feeds.”

And, Mahapatra noted, although the built-in controls will not prevent unauthorized cloud escalation, “the reason is structural rather than a knock on Perplexity’s engineering. The gate is a permission prompt, which is consent, not control. It depends on a probabilistic model correctly classifying sensitive content and correctly scoping an escalation payload, and on a user judging a request they cannot fully inspect. Both fail adversarially.”

He pointed out that, with Gmail, Drive, Slack, and GitHub connectors on a device that also holds an authorized cloud path, “this is the same connector-plus-egress combination

behind every Copilot exfiltration chain published this year. What would satisfy an enterprise security review is network-layer, not application-layer, and it does not exist in the product.”

It would need a mandatory egress proxy, DLP inspection on every escalation payload, deterministic classification rules that block defined data categories regardless of the model’s judgment, and immutable logging of what left the device, he said. “Sandboxed execution addresses code isolation and does nothing for data egress governance.”

His question to Perplexity is whether an enterprise administrator can define escalation policy centrally in a way the local model cannot override, and produce a full audit log of what crossed the boundary. “If escalation is governed by user consent and model judgment, this is a consumer product with a strong privacy story rather than an enterprise product with a compliance story,” he said, adding, “Whoever ships centrally managed escalation policy with enforceable egress inspection and tamper-evident audit will own the regulated-industry local AI market.”

Perplexity responds

In response to a request for an interview, Perplexity Communication Manager **Beejoli Shah** [<https://www.linkedin.com/in/beejoli/>] instead provided an emailed statement pushing back on the idea that data could travel to the cloud without deliberate permission. She wrote, “content in a local document can’t authorize an escalation by itself, nor can it override product controls. Escalation to the cloud requires explicit per-action approval in addition to toggling the app out of default local-only mode.”

Shah said such a data migration can only happen if the user has already toggled “allow advisor escalation” to “on” in app settings. “When that isn’t toggled on, no work can proceed to the cloud,” she said, pointing out that the product enforces restrictions on what local data and outbound actions are available to the agent.

She added that the user must also review a request in-app to send a piece of the task to the cloud, and that, she said, as illustrated in the **introductory video** [https://x.com/perplexity_ai/status/2092268362386780270/video/1], the request pop-up is the same size as prompt input. Furthermore, she said, “Escalation is only allowed once, not across the remainder of the task, or in future sessions.”

Artificial Intelligence[<https://www.computerworld.com/artificial-intelligence/>]

NEWSLETTER

The latest AI updates, straight to your inbox

News, analysis, and insights for IT leaders navigating the risks and rewards of AI. A special series from the editors of CIO, Computerworld, CSO, InfoWorld and Network World

Email Address

☐

By submitting your information, you agree to our [PRIVACY POLICY](https://foundryco.com/privacy-policy/) [https://foundryco.com/privacy-policy/].

SUBSCRIBE

About

About Us

Advertise

Contact Us

Editorial Ethics Policy

Foundry Careers

Reprints

Newsletters

Add Computerworld as a Preferred Source in Google Search

More

News

Features

Opinion

How-To

Blogs

BrandPosts

Events

Policies

Terms of Service

Privacy Policy

Cookie Policy

Copyright Notice

Member Preferences

About AdChoices

Your California Privacy Rights

Our Network

CIO

CSO

InfoWorld

Network World

Podcasts

Videos

Buyer's Guides

FOUNDRY



© 2026 FoundryCo, Inc. All Rights Reserved.