

Few Chickens and Even Fewer Eggs

Local AI is growing fast, but the usable installed hardware base is minuscule. Great applications are even thinner.

By [Hans Kuepper](#), PromptQuorum · October 23, 2026 · Data cut-off October 23, 2026 · Version 1.0

Key findings

1. In the July 2026 Steam hardware survey, 25.9% of surveyed systems had 16 GB of graphics memory and 5.4% had 24 GB. ^[S1]
2. On a 24 GB Mac in the author's tests, the system used 9.67 GB of memory right after startup, before any model was loaded. ^[S22]
3. Of nine local AI apps the author tested, eight could be scored; four of those eight rated 2 or below on core function. ^[S22]
4. Between 2.2% and 4.5% of US households or consumers pay for an AI chatbot, according to three card-data sources. The top 1% of payers account for 19.5% of consumer AI spend. ^[S11, S12, S13]
5. The list price of the 128 GB NVIDIA DGX Spark rose from \$3,999 at launch to \$4,699 in February 2026 and \$6,950 in October 2026. ^[S3, S4]
6. Sales of new laptops and desktops fell in the third quarter of 2026: IDC reports shipments down 20.1% to 62.7 million units, Omdia down 21.2% to 58.1 million. ^[S43]
7. The largest open models trail the best cloud models by about four months, but need at least 377 GB of memory (a calculation). ^[S14, S19]
8. A search for "local AI" returns 10 results in Apple's German App Store and 29 listings in Google Play, out of about 2.75 million and 2.65 million apps. ^{[S34, S35,}

9. Google searches for "local ai" peaked in late May 2026 and by mid-October were at half their level of a year earlier; the report states that this divergence from other indicators cannot be explained from the data. ^[S29]

PLAIN-LANGUAGE EDITION · BY HANS KUEPPER · 23 OCTOBER 2026 · THE PRIVATE VIEW OF THE WRITER, BASED ON HIS EXPERIENCE DEVELOPING PROMPTQUORUM, AN ECOSYSTEM AROUND LOCAL AI

Local AI is not there yet. Very few computers can run a good model, even fewer apps make it worth the effort, and the promises that surround it — cheap, private, as good as the big cloud models — do not yet give ordinary people a reason to switch. The author calls this few chickens and even fewer eggs.

The chickens: very little hardware

Nobody publishes what share of all laptops and desktop PCs can run a good local model. The closest proxy is Valve's Steam survey of gamers, who are better equipped than most people. Local AI models run mainly in the memory of the graphics card, which is separate from the computer's main memory. A PC can have 32 GB of main memory and still only 8 or 16 GB on its graphics card, and the model has to fit on the card. Even among the gamers, only 25.9% of systems had 16 GB of graphics memory in July 2026, and 5.4% had 24 GB. ^[S1] Apple's Macs work differently: one shared memory pool serves both jobs, which is why a 24 GB Mac can be tested at all, and why the Steam survey, where Macs are 1.9% of systems, says little about them. ^[S2] Memory on paper is also not memory in use: on a 24 GB Mac in the author's tests the system alone took 9.67 GB after startup. ^[S22]

NVIDIA DGX Spark (128 GB), list price in USD



Sources: NVIDIA price announcements as reported (S3, S4); the October figure is less well documented.
Increase since launch: +74% (calculation).

Hardware that can run big models is also getting more expensive, not cheaper, because memory is scarce.^[S6, S7] Sales of new laptops and desktops fell in the third quarter of 2026: IDC reports shipments down 20.1% to 62.7 million units and Omdia down 21.2% to 58.1 million, both pointing to the memory squeeze.^[S43] So the base of capable machines is small, and the entry price is rising.

What share of new machines is “AI-enabled”, and what does it mean? Gartner forecast that 31% of laptops and PCs shipped in 2025 and 54.7% in 2026 would be “AI PCs”; that forecast dates from August 2025, before the shortage (**⚠ older than 1 year**).^[S8] No measured 2026 share was found. The last measured shares found are Canalys figures of 14% in the second quarter of 2024 and 20% in the third quarter (**⚠ older than 1 year**).^[S44] “AI-enabled” is a chip definition. Canalys counts a laptop or desktop with a chipset for dedicated AI work, such as a neural processing unit (NPU); Microsoft’s Copilot+ PC requires an NPU of at least 40 trillion operations per second, 16 GB of memory and 256 GB of storage.^[S44, S45] None of the definitions found is based on running a large language model.

The eggs: even fewer apps worth using

The author tested nine local-AI apps, eight on a 24 GB Mac and one on an Android phone. Four of the eight that could be scored did their core job poorly.^[S22] The failures are often about the interface. Apps list models without saying how big they are or whether the computer can run them. One suggested a download larger than the Mac’s memory; another listed 129 GB of downloads whose main model alone (41.5 GB) is larger than the Mac’s 24 GB of memory, with no visible memory warning. Four show no remaining download time. One never says which model it installed.^[S22] The most complete app in the series, LU, shows each model’s size and whether the Mac can run it, and its coding agent built a working game; its document chat did not read a PDF, an Excel file or a Word file.^[S22]

Almost nothing is available yet. Apple’s App Store holds about 2.75 million apps and Google Play about 2.65 million.^[S36] A search for “local AI” returns 10 results in Apple’s German store and 29

listings in Google Play,^[S34, S35] out of millions. Where ratings are shown they run from 2.2 to 4.2 stars.

Much of what does exist sits on GitHub, a platform built for developers, not for ordinary users. In at least five of the author's nine test records (Page Assist, LTX Desktop, Unsloth, DiffusionBee, BoBe) GitHub is the vendor page, the README or the release source. Page Assist's README leads with a build from source that needs a terminal and a tool called Bun, and the tester's first attempt failed.^[S22] And there is a lot of it: GitHub lists 5,794 public repositories under the topic "ollama" and 9,477 under "local-llm".^[S46] Many are forks and copies of one another. In the author's words, it is a mess where nobody knows what is current, maintained or safe. For an ordinary user, an app that lives only on GitHub is effectively not available.

The promises: cheap, private, as good as the frontier

Cheap. Saving money only matters to people who spend money on AI. In the US, the most advanced market, between 2.2% and 4.5% of households or consumers pay for an AI chatbot.^[S11, S12, S13] For everyone else the cloud tool is already free, and a local app has nothing to undercut. To save on subscriptions, a buyer would first have to pay for the hardware above.

Private. Privacy is the strongest promise and the weakest sales argument. In one survey, 32% of chatbot users had told a chatbot things they withheld from friends, family or doctors, and 53% did not know or were unsure that conversations may be used for AI training.^[S20] For twenty years people have traded privacy for good, free services from Google and others, and a local app that is worse than the service will not change that. Local AI's privacy promise is mostly marketing until the app is also good.

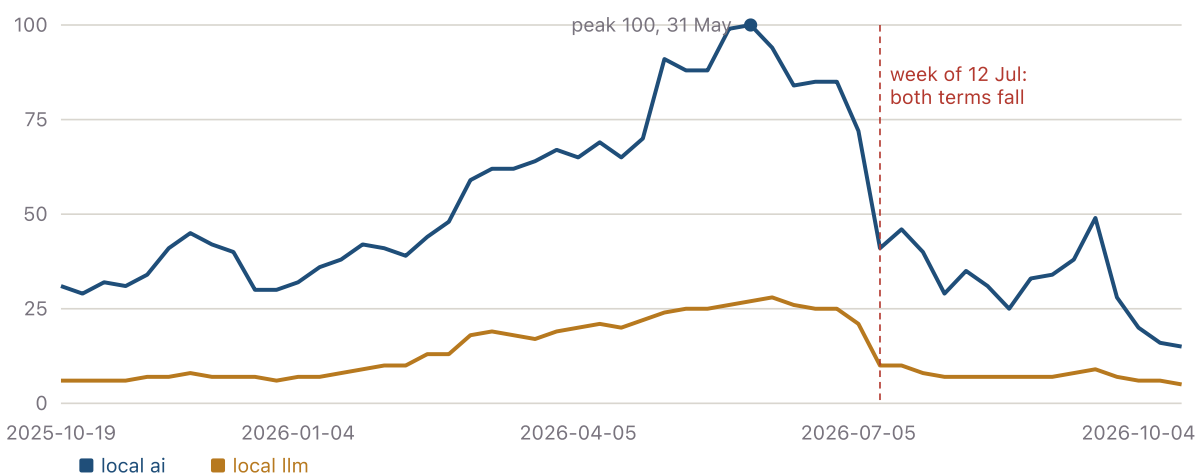
As good as the frontier. The standard claim is that open models are only months behind the best cloud models. The largest open models do trail by about four months,^[S14] but they need at least 377 GB of memory (a calculation),^[S19] far beyond any consumer machine. Models that fit one gaming graphics card trail by 6 to 12 months, according to Epoch AI, which adds that real-world lag is probably longer (a source **⚠ older than 1 year**).^[S15] For a laptop or phone there is no figure. On an Android phone in the author's tests, a 1 GB model said a sitting head of state had been assassinated, which a third-party source contradicts.^[S22] The months-behind figure is a marketing argument. Users expect frontier-level results from these apps and do not get them: the interface is complicated, the setup is complicated, and the results are poor or do not arrive at all. In the author's tests one app stopped at setup before producing anything, one ran out of memory before the first picture, and one returned a gray square after the wrong model was chosen.^[S22]

The money: very little invested

Capital follows markets, and it has not arrived. Ollama, the software that runs models, raised \$65 million in July 2026.^[S17] For the apps on top, web searches found only an old round for GPT4All's maker (2023) and a small note for AnythingLLM's, and nothing for five other named apps.^[S27, S28] Venture rounds are normally public, so the author reads this as very little investment. One app he is in contact with has just found an investor (not named).

Not there yet

Google Trends, worldwide, weekly: "local ai" and "local llm" (index 0–100)



Source: Google Trends CSV supplied by the author (S29). Index is relative: 100 = highest weekly value of either term in the window. Only 12 months are available, so seasonality cannot be seen.

Google searches for “local ai” rose to a peak in late May 2026 and by mid-October were at half their level of a year earlier.^[S29] The drop was a step in one week of July. **What is going on cannot be explained from the data.** Other indicators point up while Google Trends points down: Ollama says its developer base doubled between January and July,^[S17] and the author reports that other trends keep rising (not verified here for August to October).

Local AI is growing among developers. For ordinary people the case to switch is not yet there: too little hardware, too few good apps, promises that do not hold in use, and little capital to change that.

Summary: just the beginning

Local AI is just starting. In the developer scene it is well known: Ollama's developer base doubled in six months, and Chinese labs released the largest open model almost every month of 2026.

[S17, S19] In the general public it is barely known at all. Most people do not even know what is possible.

And when they try? First they run into the hardware. Their laptop or PC is not good enough, and even among gaming machines only a small share has the graphics memory a good model needs. [S1] Then they start using the apps, and the apps mostly underperform compared with the frontier models they know from the cloud. [S22]

So this market is dominated, as every new technology market is, by early adopters and enthusiasts. That is changing fast: models are getting more capable for their size. [S37]

The same pattern usually holds across countries. New technologies tend to start in the advanced economies and spread to the rest later. [S31] The author's impression is that the United States and China are furthest along with local AI, since most AI models are built there and the largest open models come from Chinese labs, [S19] while most other countries have barely started. No country-level figures for local AI exist; the only country data on AI use in general shows a wide gap between countries. [S30] The author expects the wave to become global.

Hardware will follow, at a price. Computers will become more and more AI-enabled, but how capable they are depends on what they cost. That is the crux. For devices with 64 GB of memory and more to sell in volume, prices need to fall substantially, and the forecasts do not agree that they will. Counterpoint expects new memory capacity to bring prices down by 2028, [S38] but TrendForce expects consumer memory prices to keep rising next year [S38] and IDC sees the shortage lasting into 2027. [S7]

The apps will follow too. New app releases rose 60% in the first quarter of 2026 against a year earlier, and 104% in April to date, according to Appfigures. [S41] AI coding tools are behind the surge, in the app stores and on GitHub. Creativity is moving to individuals, who can do today what they could not do years ago. Individuals can work with local AI and build their own apps, unique and tailored to their needs, and the cost of building an app has fallen substantially.

If local AI follows the path of other technologies, hardware is likely to become more capable and affordable, apps are likely to get better, and AI coding tools are likely to speed up app development. The author expects the future to be hybrid: a growing share of AI will run on the user's own device where privacy, cost or convenience justify it, and the rest will stay in the cloud.

We are just not there yet. Today there are few chickens and even fewer eggs.

The writer's view. This report is the writer's impression of the local-AI market. It rests on his experience building PromptQuorum, on nine apps he tested himself, and on the public figures cited. The conclusions, that hardware and apps are both still thin and that both will improve, are his own assessment. They are not forecasts from the sources.

Key takeaway. Few chickens, even fewer eggs: today little hardware can run local AI, few apps are worth using, and the promises of low cost, privacy and frontier quality do not yet hold for ordinary users. But models, hardware and apps are all moving fast, so the gap is closing. We are just not there yet.

Free to quote and reuse

You may quote any sentence or figure from this report, and reuse its charts and tables, in articles, newsletters, talks and social posts, without asking. This applies to the PDFs and CSV files as well. It is licensed under [CC BY 4.0](#). What we ask: name the author and PromptQuorum, **link to this page**, and say if you changed anything.

This covers our own text, figures, charts and tables. The third-party sources behind the numbers (the [S...] references) stay under their owners' terms; cite those at the source. Please quote figures with their source label and the data cut-off date (23 October 2026), because the numbers change with each report.

LINK TO THIS REPORT

<https://www.promptquorum.com/about/state-of-local-ai/2026-q4>

HTML LINK FOR YOUR PAGE

```
<a href="https://www.promptquorum.com/about/state-of-local-ai/2026-q4">The State of Local AI, Report 01 (Q4 2026) - PromptQuorum</a>
```

Questions, a quote check or a corrected number: hello@promptquorum.com. If you publish something based on this report, send us the link and we will list it here.

Cite this report

APA

Kuepper, H. (2026, October 23). Few chickens and even fewer eggs (The State of Local AI, Report 01, Q4 2026). PromptQuorum.

<https://www.promptquorum.com/about/state-of-local-ai/2026-q4>

CHICAGO

Kuepper, Hans. "Few Chickens and Even Fewer Eggs." The State of Local AI, Report 01 (Q4 2026). PromptQuorum, October 23, 2026.

<https://www.promptquorum.com/about/state-of-local-ai/2026-q4>.

PLAIN TEXT

Kuepper, H. (2026). Few Chickens and Even Fewer Eggs. The State of Local AI, Report 01 (Q4 2026). PromptQuorum.

<https://www.promptquorum.com/about/state-of-local-ai/2026-q4>

About this report

The author founded PromptQuorum, a content site and directory about local AI apps, and tested apps that are listed there. Each hands-on test record, with evidence labels and the tester's own errors, is linked from the evidence file. No payment or sponsorship is recorded in the test notes.

Hans Kuepper is the founder of PromptQuorum, based in Frankfurt, with more than 30 years of consulting and general management experience.

The method, labels, assumptions and sources are in the full evidence section above, which also exists as a separate page.

Press contact: hello@promptquorum.com

Corrections and updates

- October 23, 2026: Version 1.0: first publication.