

# Evidence file: Few Chickens and Even Fewer Eggs

The same findings with labels, calculations, assumption boxes, a claim register and sources. Cite from here.

The State of Local AI · Report 01 · Q4 2026 · [Read the report](#)

## Contents

0. [How to read this report](#)
1. [Summary of findings](#)
2. [Thesis and definitions](#)
3. [Few chickens: the hardware base](#)
4. [Even fewer eggs: applications and willingness to pay](#)
5. [Local models versus frontier models](#)
6. [Demand signals and what is changing](#)
7. [Trade and supply-chain implications \(EU, China\)](#)
8. [Outlook, with assumptions labelled](#)
9. [Claim register](#)
10. [Sources](#)

---

## 0. How to read this report

Every statement carries one of the labels below. A statement without a label is connective text and contains no factual claim. Where a number comes from a secondary report rather than the original publisher, the source list says so. Sources older than twelve months are marked **⚠ older than 1 year** in Section 11.

**FACT** Reported by a named third-party source (citation given).

**AUTHOR-MEASURED** Measured or observed by the author in PromptQuorum hands-on tests; evidence labels from the tests are kept.

**CALCULATION** Arithmetic performed on cited figures; the calculation is shown.

**THIRD-PARTY FORECAST** A third-party projection, not an observation.

**ASSUMPTION** An inference. The box names **who** makes it and **what it rests on**.

**UNVERIFIED** Stated by the author; no supporting data found yet.

## Disclosure and method limits

- The author founded PromptQuorum, which operates a content site, a directory of local-AI applications and a desktop application for multi-model prompting. Applications in the directory receive traffic from the site. No payment, sponsorship or vendor contact is recorded in the test notes.
- The hands-on sample is **nine tests** by **one tester**, eight on one 24 GB Apple Mac (recorded as M3 Pro in earlier records; the LU record corrects its own “M3” note to M4, so the chip is recorded inconsistently) and one on an Android phone (12 GB). It is a convenience sample, not a random sample of the app market. The tester founded PromptQuorum, which lists the tested apps; for LU the developer supplied a founder statement and exchanged messages with the tester before and after the test (disclosure in the test record). No payment is recorded.
- Several test records list the tester's own errors (for example, choosing the wrong model type, or judging features that were visible but not evaluated). These are kept in the underlying test pages and are summarised here where they affect a finding.
- Consumer-spending data in this report is almost entirely US-only. No comparable EU household data was found.

---

## 1. Summary of findings

The report tests one hypothesis: that local AI adoption is held back by a small installed base of capable machines (the chickens) and a thin supply of applications that justify using them (the eggs). The table gives the result per component, with the strength of evidence.

| #  | Finding                                                                                                                                                                                                                                                                                                                                                                                                                | Evidence                                   | Strength                                    |
|----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------|---------------------------------------------|
| 1  | Capable hardware is a minority of even the best-case population. In the July 2026 Steam survey, 51.4% of opted-in PC gamers had 10 GB VRAM or more, 5.4% had 24 GB. <sup>[S1]</sup>                                                                                                                                                                                                                                    | FACT                                       | Medium: gamer sample, opt-in                |
| 2  | Memory budget is consumed by the system before any model loads. A 24 GB Mac used 9.67 GB right after startup and 12.92 GB with ordinary apps open. <sup>[S22]</sup>                                                                                                                                                                                                                                                    | AUTHOR-MEASURED                            | Medium: one machine                         |
| 3  | Entry prices rose in 2026 while the usual pattern is falling prices. NVIDIA DGX Spark (128 GB): \$3,999 at launch, \$4,699 (Feb), \$6,950 (Oct). 16 Gb DRAM: +49% from Q1 to Q2 2026. <sup>[S3, S4, S6]</sup>                                                                                                                                                                                                          | FACT<br>CALCULATION                        | High for Spark (Feb), medium for Oct figure |
| 4  | Few people pay. In the US, the most advanced market, 2.2% to 4.5% of households or consumers pay for a chatbot subscription (PNC 2.2%, Bank of America about 3%, YipitData/a16z 4.5%). No comparable EU data was found. Why it matters: cost saving is the main argument for local AI, and it only works if people spend money on AI today. <sup>[S11, S12, S13]</sup>                                                 | FACT                                       | High: three independent card datasets       |
| 5  | Result of the author's own hands-on tests (9 apps; eight on a 24 GB Mac, one on an Android phone): 4 of 8 scored apps rate 2 or below on core function. Test pages are linked in Section 4.2. This is the author's test series, not a market sample. <sup>[S22]</sup>                                                                                                                                                  | AUTHOR-MEASURED                            | Low-medium: n=9, one tester                 |
| 5a | User interfaces are a recurring problem in the tests: six of seven scored apps rate 3 or below on onboarding (only LU rates 4); the same patterns (no model size or hardware fit, no remaining download time, unsuitable models offered without a memory warning, state not explained) recur across vendors. Detail in Section 4.2. <sup>[S22]</sup>                                                                   | AUTHOR-MEASURED                            | Low-medium: n=9, one tester                 |
| 6  | The "6–12 months behind" statement is correct only for models small enough for a single consumer GPU (24–32 GB VRAM). The largest open models lag by ~4 months but do not run on consumer hardware. The author regards the lag figure as a marketing argument: the models it measures are not the ones users can run. <sup>[S14, S15]</sup>                                                                            | FACT<br>ASSUMPTION                         | High (Epoch AI); assessment is the author's |
| 7  | Supply is moving in the direction of the thesis: Ollama raised \$65M in July 2026; Gartner forecasts AI PCs at 55% of 2026 shipments, though the definition is NPU-based, not LLM-capable. <sup>[S8, S17]</sup>                                                                                                                                                                                                        | FACT<br>THIRD-PARTY FORECAST               | Medium                                      |
| 8  | Speed: open models gain capability per parameter fast (density doubling about every 3.5 months to Apr 2025), new app releases are up 60–104% year on year, and memory supply is forecast to expand from 2027–2028 (forecasts disagree). Author's view: if local AI follows the general technology path, hardware and apps are likely to improve over the coming years; today not there yet. <sup>[S37, S38, S41]</sup> | FACT<br>THIRD-PARTY FORECAST<br>ASSUMPTION | Medium: trends and forecasts                |

**Key takeaway.** The data supports the "few chickens" half of the thesis for consumer hardware, and a qualified version of the "few eggs" half. It does not yet support statements about search interest, country-level adoption or app counts, which are listed as open items.

---

## 2. Thesis and definitions

### 2.1 The thesis

The classic chicken-and-egg problem asks which comes first. For local AI the problem is different: both sides are scarce at the same time. Few computers can run capable local models, and few applications exist that give a non-technical user a reason to buy such a computer.

### 2.2 Who the report is about

The reference user is a non-technical user who expects to download an app, understand what it does, optionally add a local model through Ollama or a direct download, and use it without a terminal, Python, or Linux. Developers and enterprise IT, who are today's main local-AI users, are discussed only as a contrast.

**ASSUMPTION** **Made by:** Author (Hans Kuepper)

Enterprises and developers are the early market for local AI hardware; private end users follow later.

**Based on:** Author's market observation. Supporting data in this report: coverage describes Ollama as a developer platform, and DGX Spark is positioned as a developer system. No consumer-versus-enterprise split of local-AI users was found.

### 2.3 Definitions used

| Term              | Meaning in this report                                                                                                 |
|-------------------|------------------------------------------------------------------------------------------------------------------------|
| Local AI          | A model runs on the user's own device. Hybrid products that send some requests to a cloud service are noted as hybrid. |
| Open-weight model | A model whose weights can be downloaded. Not necessarily open source in the licence sense.                             |

| Term            | Meaning in this report                                                                                                                                                                                                                                                                                                  |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Frontier model  | The most capable closed or open model at a given date, as measured by a stated index (here: Epoch Capabilities Index).                                                                                                                                                                                                  |
| Capable machine | <b>Not defined by a single threshold in sources.</b> The report uses measured memory data from the author's tests and states the model size it applies to.                                                                                                                                                              |
| AI PC           | A laptop or desktop with a chip for on-device AI (an NPU). Definitions differ: Canalys counts any chipset for dedicated AI workloads; Microsoft's Copilot+ PC requires an NPU of 40 TOPS or more, 16 GB RAM and 256 GB storage. <sup>[S44, S45]</sup> This does <b>not</b> mean it can run a large language model well. |

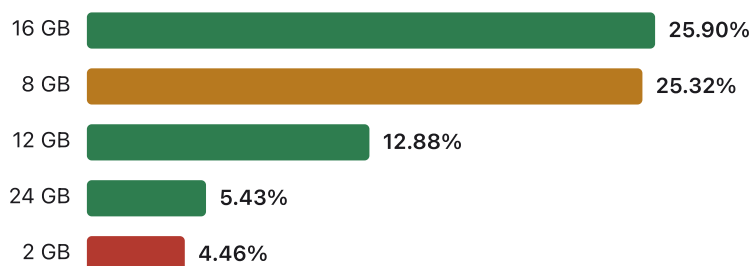
## 3. Few chickens: the hardware base

### 3.1 Best available data: the Steam Hardware Survey

**Two different memories.** A computer has **system memory (RAM)**, which every program shares, and a graphics card has its own **graphics memory (VRAM)**. A local AI model has to fit into one place to run at a usable speed, on a PC mostly the graphics card. This report therefore counts graphics memory (VRAM) only. Main-memory sizes are left out because they do not decide whether a PC can run a model: a PC can have 32 GB of RAM and 8 GB of VRAM. Macs and some other designs share one pool, so for them the single memory figure is the one that counts.

No public dataset reports the share of all personal computers that can run a given local model. The closest public proxy is Valve's monthly Steam survey. It is opt-in, covers gamers, and is therefore likely to overstate capability compared with the general PC population.

### Steam survey, July 2026: share of surveyed systems by GPU memory (VRAM)



Source: Steam July 2026 as reported by tBreak, Hardware Busters, Club386. Listed buckets total 74.0%; the remainder is other sizes. 10 GB or more: 51.42% (tBreak). The default Steam chart combines Windows, macOS and Linux; on **Windows only**, 8 GB remains the most common at 33.20%. September 2026: 16 GB 27.21%, 8 GB 26.71%.

[S1, S2]

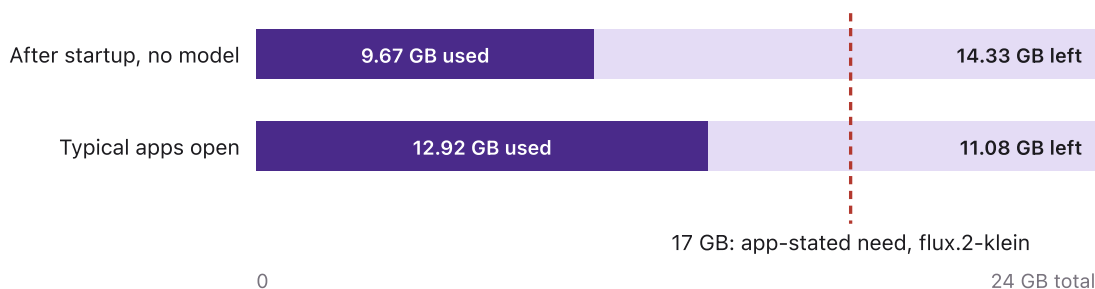
**FACT** The July 2026 survey shows 16 GB VRAM as the largest single bucket for the first time (25.90%).<sup>[S1]</sup> **FACT** Only 5.43% of respondents had 24 GB of VRAM, and a 32 GB bucket was not found in the retrieved reports. **FACT** macOS accounted for 1.92% of Steam systems in September 2026, so the survey says little about Mac users.<sup>[S2]</sup>

**Key takeaway.** Among opted-in gamers, the 16–32 GB range is now common, but 24 GB or more of VRAM is a minority. For the general population of laptops and PCs no comparable data exists, and the true share is probably lower (assumption above).

## 3.2 The system uses memory before the model does

A model size that fits in total memory does not fit in *free* memory. The author's tests give measured values on a 24 GB Apple Mac.

## 24 GB Mac: memory in use vs. memory a 17 GB image model needs



Source: author's Unsloth test, Activity Monitor readings (measured) and the app's own memory message (vendor-stated need). "GB left" = 24 minus used (calculation). Readings were taken at different moments in the test.

**AUTHOR-MEASURED** Activity Monitor showed 9.67 GB used right after startup and 12.92 GB used in a first capture with a browser, office and chat applications open. <sup>[S22]</sup>

**AUTHOR-MEASURED** The Unsloth app refused a 17 GB image model; it counted about 2 GB usable (7 GB free) before a reboot and about 6 GB usable (11 GB free) after. <sup>[S22]</sup>

**CALCULATION**  $24 - 9.67 = 14.33$  GB left after startup, which is below the 17 GB the app stated for that model.

**AUTHOR-MEASURED** A 27B-parameter model in 4-bit format ran in chat at 11.0 tokens per second on the same machine, but reading a web page and an event search both ended in "Ran out of memory" (at 34.4k context). <sup>[S22]</sup>

**ASSUMPTION** **Made by:** Author (Hans Kuepper)

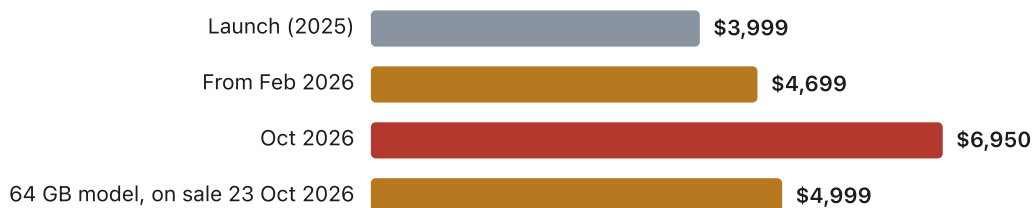
Even 32 GB of VRAM may not be enough for a normal user who keeps an editor and other applications open next to the model.

**Based on:** Author's reasoning from the 24 GB measurements above. No test on a 32 GB machine exists in the data; the statement is an extrapolation.

**AUTHOR-MEASURED** Guides often say macOS uses about 4 GB. In the author's normal use of a 24 GB Mac the system used 9.67 GB right after startup. The guide figure describes a bare system; the measured figure is what an ordinary user experiences, and it is the one that decides how much memory is left for a model. The figures are from a normal working machine, not a laboratory set-up. <sup>[S22]</sup>

### 3.3 The price of a capable machine

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



Sources: NVIDIA forum (Feb 2026, primary); Implicator.ai (Oct 2026, secondary, NVIDIA page not retrieved). Partner GB10 systems were reported at \$7,000–9,000 (Tom's Hardware via Implicator.ai).

[S3, S4]

**CALCULATION**  $\$6,950 / \$3,999 = 1.74$ , i.e. +74% since launch. The February increase was +\$700 (+18%); the October increase was +\$2,251 (+48%). NVIDIA attributed both to memory supply constraints. [S3, S4]

**FACT** Component pricing shows the same direction: the average 16 Gb DRAM price rose from \$19.2 to \$28.5 between Q1 and Q2 2026. [S6] **CALCULATION**  $28.5 / 19.2 = 1.48$ , i.e. +48% to +49%. One outlet reports +89% for the same data; the underlying figures do not support it. [S6]

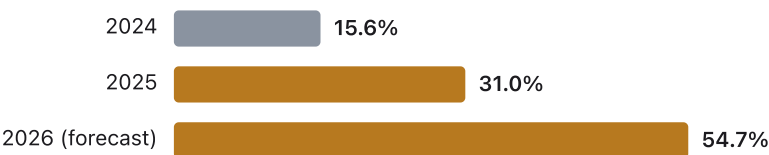
**FACT** IDC expects the DRAM/NAND shortage to last into 2027 and warns of up to 8.9% fewer PC shipments in the worst case; Dell and Lenovo announced 15–20% price increases. [S7]

**FACT** Counter-example: Apple raised entry prices of the M5 MacBook Air (\$100) and the 14-inch MacBook Pro with M5 Pro (\$200) in March 2026 but left RAM-upgrade pricing unchanged (24 to 48 GB on the 14-inch: \$400, as in the previous generation). [S5] Later reports of further Apple price rises exist but come from a single observer and were not verified.

**Key takeaway.** The usual pattern of falling prices for new hardware did not hold in 2026 for memory-heavy machines. The exception is Apple, whose RAM upgrade pricing was unchanged at the March 2026 launch.

## 3.4 The installed base: forecast versus capability

### Gartner forecast: AI PCs as share of worldwide PC shipments



Gartner, August 2025 (older than 1 year). Shipments: 77.8M in 2025, 143M forecast for 2026. 2024: about 38M (IT Reseller, citing Gartner). Forecast pre-dates the 2026 memory shortage.

[S8]

**CALCULATION** 77.8M + 143M = 220.8M AI PCs shipped in 2025–2026 on Gartner's forecast; with about 38M in 2024 the three-year sum is about 259M. This is consistent with the statement that a base of "hundreds of millions" of AI-labelled devices will exist well before 2030.

### NEW LAPTOPS AND PCS: AI-ENABLED SHARE, AND WHAT THE LABEL MEANS

| Source             | Period  | AI-enabled share of shipments                                                                          | Definition used                                                                                                      | Date flag           |
|--------------------|---------|--------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|---------------------|
| Canalys            | Q2 2024 | 14% (8.8M units)                                                                                       | Desktop or notebook with a chipset for dedicated AI workloads, such as an NPU <sup>[S44]</sup>                       | ⚠ older than 1 year |
| Canalys            | Q3 2024 | 20% (13.3M units)                                                                                      | Same <sup>[S44]</sup>                                                                                                | ⚠ older than 1 year |
| Gartner (forecast) | 2025    | 31.0% (77.8M units)                                                                                    | Not formally defined on the retrieved page; tied to on-device AI features and the processor platform <sup>[S8]</sup> | ⚠ older than 1 year |
| Gartner (forecast) | 2026    | 54.7% (143.1M units)                                                                                   | Same; forecast made in August 2025, before the 2026 memory shortage <sup>[S8]</sup>                                  | ⚠ older than 1 year |
| IDC / Omdia        | Q3 2026 | Total shipments -20.1% / -21.2% year on year (62.7M / 58.1M units); no AI share given <sup>[S43]</sup> | -                                                                                                                    | current             |

**CALCULATION** IDC: 62.7M / 78.5M = 0.80, i.e. -20.1%. Gartner 2026: 143.1M AI PCs at 54.7% implies total 2026 shipments of about 262M, which was the forecast before the Q3 fall.

**FACT** Microsoft's Copilot+ PC floor of 16 GB RAM is the only memory requirement found in any AI-PC definition; none is based on running a large language model.<sup>[S45]</sup>

## 4. Even fewer eggs: applications and willingness to pay

### 4.1 How many people pay for AI at all

The author's starting point was that about 2–3% of US households pay for AI, so a "free" local alternative has limited appeal. Three independent US card datasets confirm a low single-digit share. This report uses only that range.

#### US: share paying for AI, by source



PNC and BofA: payment data of their own customers. a16z/YipitData: card panel, paid personal ChatGPT, Gemini or Claude subscriptions only (up from 2.1% a year earlier). The datasets use different bases (households vs. consumers), so the range 2.2–4.5% is reported rather than one number.

[S11, S12, S13]

**FACT** Card data from PNC (2.2%, May 2026), Bank of America (~3%) and YipitData via a16z (4.5%, up from 2.1% a year earlier) agree on a low single-digit share. [S11, S12, S13] **FACT**

The top 1% of payers account for 19.5% of observed consumer AI spend, averaging \$903 per month. [S13]

### 4.2 What the hands-on tests show

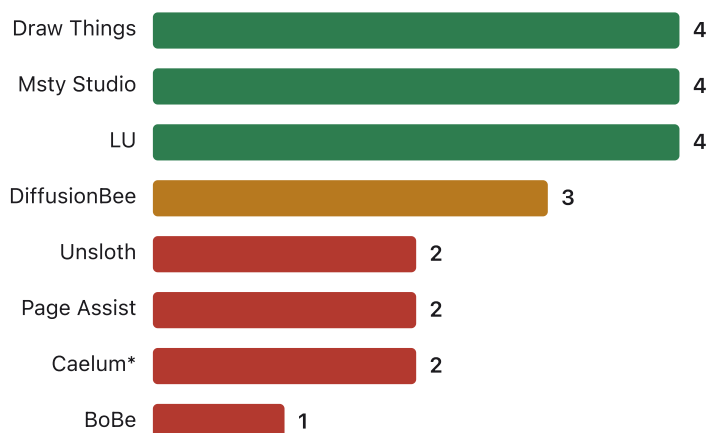
**These are results of the author's own hands-on work, not a market survey.** Nine applications were tested between 6 and 24 October 2026, eight on a 24 GB Apple Mac and one on a 12 GB Android phone. Each app name links to the full test page with evidence labels, scores and the tester's own errors. [S22]

| App                         | Type             | Status    | Core function | Download time shown | Model size / fit shown | Tester's verdict                                      |
|-----------------------------|------------------|-----------|---------------|---------------------|------------------------|-------------------------------------------------------|
| <a href="#">Draw Things</a> | Image generation | Completed | 4             | No (percent, MiB)   | No                     | Not continuing; could not find how to revise an image |

| App                            | Type                                  | Status        | Core function | Download time shown                    | Model size / fit shown             | Tester's verdict                                                               |
|--------------------------------|---------------------------------------|---------------|---------------|----------------------------------------|------------------------------------|--------------------------------------------------------------------------------|
| <u>DiffusionBee</u>            | Image generation                      | Completed     | 3             | No (percent only; one timeout)         | No                                 | Best local image app tried; editing mostly failed                              |
| <u>Msty Studio</u>             | Chat workspace                        | Partial       | 4             | Not recorded                           | Yes (hardware-fit hint)            | No verdict recorded                                                            |
| <u>Unslow</u>                  | Run / train models                    | Partial       | 2             | Yes in hub; no progress at model start | Parameters, not GB                 | Not for this machine, for now                                                  |
| <u>Page Assist</u>             | Browser sidebar                       | Abandoned     | 2             | Not recorded                           | n/a                                | Chat only; uninstalling                                                        |
| <u>BoBe (alpha)</u>            | Screen-aware assistant                | Abandoned     | 1             | No                                     | Yes (RAM per model)                | Concept liked, implementation missing                                          |
| <u>Locally Uncensored (LU)</u> | All-in-one: chat, image, coding agent | Partial       | 4             | No (percent and MB/s only)             | Yes (size, memory need, fit label) | Most complete single local app seen so far, one to watch; document chat failed |
| <u>Caelum (Android)</u>        | Phone chat app                        | Completed     | 2*            | Yes (ETA shown)                        | No model name given                | Would not keep                                                                 |
| <u>LTX Desktop</u>             | Video generation                      | Not completed | n/a           | Not recorded                           | Downloads 129.1 GB listed          | Stopped at setup; below vendor memory floor                                    |

\*Score proposed from the evidence; the tester has to confirm. Core-function scores are the tester's dimension scores; the scale maximum is not stated in the data. n/a = not tested.

## Core-function score per tested app (higher is better)



Axis set to 0–5 (scale maximum is not stated in the test data). LTX Desktop is not scored.

[S22]

## USER INTERFACE: WHAT THE TESTS RECORDED

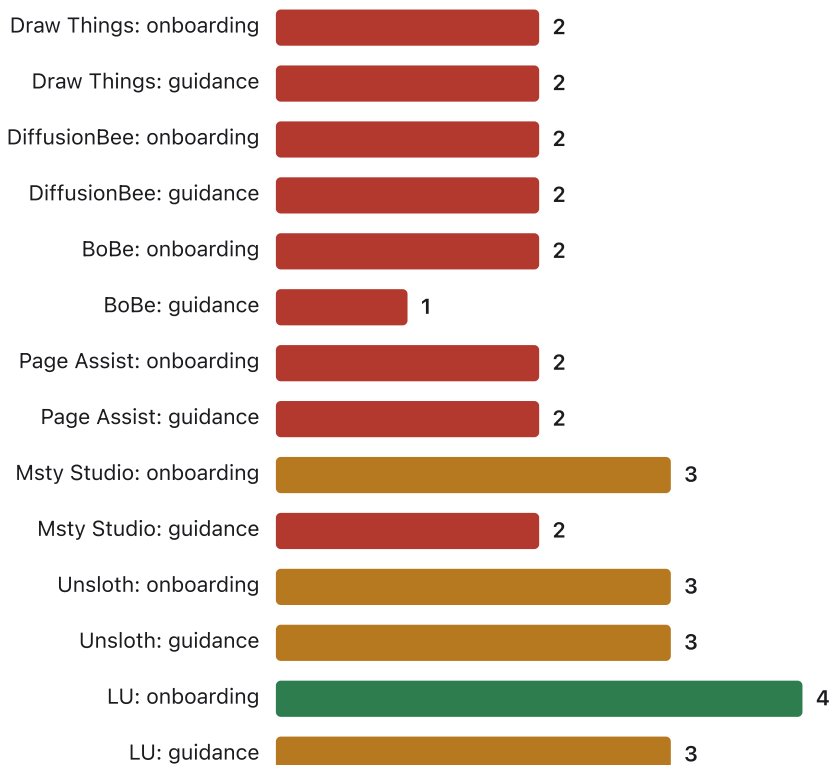
**AUTHOR-MEASURED** Each row lists interface problems from the test findings (finding IDs and evidence label as recorded; "observed" unless stated), the strengths recorded, and any misreading by the tester that touches the same point. Interface problems are not scored separately in the tests; the table is a reading of the findings.<sup>[S22]</sup>

| App                 | Interface problems recorded                                                                                                                                                                                                                                                                                                                                                                      | Strengths recorded                                                      | Related tester misreading                                                                                                     |
|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| <u>Draw Things</u>  | Model list: no size, task or hardware fit (DT-01, High). Welcome dialog suggests FLUX.1 with no memory caveat; downloads read as 25–57 GB on a 24 GB Mac (DT-02, High). No guard against a video-restoration model for a text prompt, result a gray square (DT-03, High). Download shows no remaining time (DT-04); app unusable during download (DT-05). Controls unclear (DT-11, tester view). | Install and image speed once a suitable model ran (scores 4).           | Wrote that the list has no search (a search field exists); judged iteration without evaluating Version History or Edit tools. |
| <u>DiffusionBee</u> | Model cards: no size, task or fit (DB-01, High). FLUX cards say the app must be updated although newly installed (DB-02, High). Percentage only, no size or time (DB-04); a timed-out download offered only 'Download Again' (DB-05). Input Strength 100 crashes the app, per notes without screenshot (DB-18, High). App does not explain what a model is (DB-21, tester view).                 | Queue for several pictures; 'Actions' button on results (DB-08, DB-24). | Tester suspects he mishandled Input Strength (assumption).                                                                    |

| App                | Interface problems recorded                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Strengths recorded                                                                                                                                                             | Related tester misreading                                                                                                     |
|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| <u>BoBe</u>        | Screen-recording permission step unclear (BB-06). In-app assistant did not name the macOS setting after two questions (BB-07, High). No remaining time (BB-05). Announced 11 GB, downloaded 4.9 and 5.7 GB, unexplained (BB-04). App restarted when deleted (BB-10, High).                                                                                                                                                                                                                                                            | Clear three-model dialog with RAM needs (BB-03); 312 MB install, one approval (BB-01).                                                                                         | Wrote there is no clear way to set up screen access; an 'Open System Settings' link is shown.                                 |
| <u>Page Assist</u> | Toolbar icon opens a full-screen page, sidebar found by right-click only after help from ChatGPT (PA-04, High). Welcome-video shortcut did not work (PA-05). First side-panel question showed only a cursor (PA-08). Two chats returned empty 'Thought for Oms' replies (PA-07, High). 'Add MCP server' does nothing visible (PA-15).                                                                                                                                                                                                 | Word count of a page worked (PA-09).                                                                                                                                           | Missed the store link in the README; sidebar route miss is shared with the welcome video.                                     |
| <u>Msty Studio</u> | First screens do not say what the app is for (MS-01, tester view). Image-only and cloud models sit in the same picker as chat models (MS-05). File upload looped between knowledge base and chat (MS-08, High, tester view). No local image model selectable in Media Studio (MS-10, High).                                                                                                                                                                                                                                           | Existing Ollama models found at once (MS-03); hardware-fit hint on model cards (MS-11).                                                                                        | Believed the Ollama connection had failed; it had worked. Wrote the app does not say 'Ollama'; a screenshot shows the header. |
| <u>Unsloth</u>     | Picker shows parameter counts, not gigabytes (UN-04). Model start shows a spinner with no percentage or time for two to three minutes or more (UN-23). 13-step dependency install of 4–5 minutes with no estimate (UN-02).                                                                                                                                                                                                                                                                                                            | Hardware readout, download time left, size warnings (UN-09, UN-11); refusal message explains cause and names an override (UN-20); context meter and tokens per second (UN-24). | Read parameter counts as sizes; wrote there is no token counter (one is shown).                                               |
| <u>LTX Desktop</u> | Setup listed 129.1 GB of downloads (five files). The main model, LTX 2.5, is 41.5 GB, larger than the Mac's 24 GB of memory, and no memory warning was visible in the screenshots (LTX-03, High, observed). Download size is not runtime memory (the test record's own assumption); the vendor README lists 32 GiB of RAM as the floor for local generation (LTX-01, vendor claim). The 129.1 GB total also exceeds the 32.9 GB of free disk; whether the app blocks was not tested (LTX-04). Hugging Face sign-in required (LTX-05). | Smooth first launch (LTX-08, tester view).                                                                                                                                     | Did not check requirements before installing.                                                                                 |

| App                            | Interface problems recorded                                                                                                                                                                                                                                                                                                                                            | Strengths recorded                                                                                                                                                                                                                                     | Related tester misreading                                                                                                                                                                        |
|--------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <u>Locally Uncensored (LU)</u> | PDF (872 KB) and Excel file attached with the paperclip not read; Word file listed as 20 chunks yet reported unreadable (LU-10, LU-11, LU-13; High for PDF and Word). Download dialogs show percentage and MB/s but no time left (LU-15). Sampling settings (Temperature, Top P, Top K) unexplained (LU-23, tester view). Troubleshoot window would not close (LU-24). | Three-step setup, existing models found (LU-04); size, memory need and fit label on image-model cards (LU-17); context meter (LU-09); built-in explainer (LU-20); coding agent built a working game (LU-29); A/B compare and benchmark (LU-27, LU-28). | Expected the engine tick to toggle; tested .xlsx, which the Docs zone does not list; judged image quality on one 512 px starter model; did not record whether the embedding model was installed. |
| <u>Caelum (Android)</u>        | Never says which model it downloaded (CAE-10). Web-search button has two blue shades, no label (CAE-06). App does not stress that the AI runs on the phone (CAE-14, tester view). Unclear when the web interface is on (CAE-15).                                                                                                                                       | Download with progress bar, size and ETA (CAE-01); edit, regenerate, copy, read-aloud buttons (CAE-16).                                                                                                                                                | -                                                                                                                                                                                                |

### Onboarding and guidance scores per app (higher is better)



Source: dimension scores in the author's hands-on tests (S22). Caelum and LTX Desktop were not scored on these dimensions; LU is scored. Axis 0–5 (scale maximum is not stated in the test data).

[S22]

## PATTERNS ACROSS THE TESTS

| Pattern                                                               | Apps showing it                                                                                                  | Apps showing the opposite                                                                                   | Not recorded                       |
|-----------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|------------------------------------|
| No remaining time on model downloads                                  | Draw Things, DiffusionBee, BoBe, LU (4)                                                                          | Unsloth (time left), Caelum (ETA) (2)                                                                       | Page Assist, Msty, LTX Desktop (3) |
| Model list without size, task or hardware fit                         | Draw Things, DiffusionBee (2); Unsloth shows parameter counts only (partial)                                     | BoBe (RAM needs), Msty (hardware-fit hint), LTX Desktop (sizes in setup), LU (size, memory need, fit label) | Page Assist, Caelum                |
| Unsuitable model or download offered without a visible memory warning | Draw Things (FLUX.1 suggestion, download sizes above 24 GB), LTX Desktop (41.5 GB main model on a 24 GB Mac) (2) | BoBe (RAM stated per model), Unsloth (several warnings), LU ('Too big for this Mac' labels)                 | others                             |
| App does not explain its own state or the model used                  | BoBe (assistant), Page Assist (empty replies), Caelum (model unnamed), DiffusionBee (what a model is)            | Unsloth (refusal explains cause)                                                                            | others                             |

### ASSUMPTION **Made by:** Author (Hans Kuepper)

Very few local-AI apps have a user interface in which everything is developed, and for non-technical users this is now a bottleneck as large as hardware.

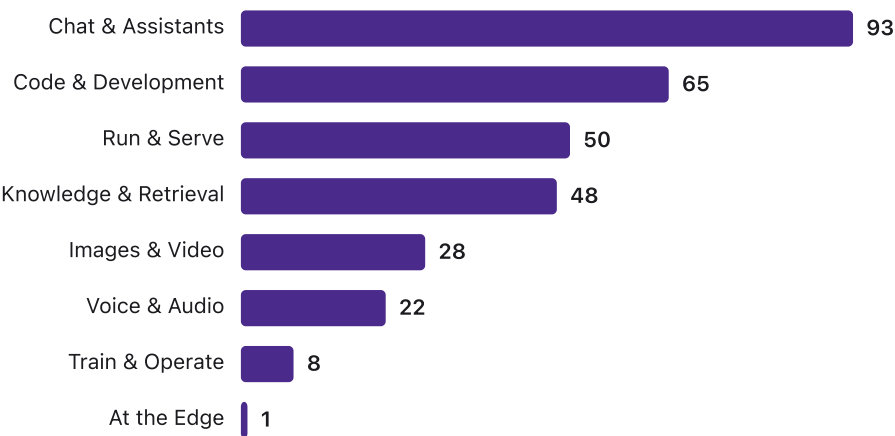
**Based on:** The tests above: six of seven scored apps rated 3 or below on onboarding (only LU rated 4) and only two rated 3 on guidance; the same four patterns recur across different vendors. Limits:  $n=9$ , one tester, convenience sample; at least five records list a related misreading by the tester, so counts are an upper bound; whether a missed control is an interface fault is a judgement. No data exists on how a typical non-technical user would fare. Counter-evidence: Unsloth's interface was rated polished by the tester, LU's first-run setup and in-app explainers were rated strong, and BoBe, Msty and Caelum have clear elements recorded.

**Key takeaway.** The tests support a pattern (missing model information, missing time estimates in some apps, gaps in editing) but not precise percentages. With nine tests the report can state counts, not rates.

## 4.3 The size and shape of the app market

No app store for local-AI software exists; applications are spread across GitHub, vendor websites and the mobile stores. The author's own directory gives one structured view.

PromptQuorum directory: listings per category group, 23 Oct 2026



Source: directory category counts (author's own directory). Group counts may include an app in more than one group; the sum of 315 is therefore an upper bound on distinct apps. General chat clients: 74 listings (23% of 315). Contract review: 1 listing. Curated directory: not a census of the market.

[S23]

**AUTHOR-MEASURED** Within the directory, 74 of up to 315 listings are general chat clients, which is consistent with the observation that many local applications are chat front ends on top of a downloaded model. [S23]

STORE SEARCHES, 23 OCTOBER 2026

| Store                          | Query    | Result                                             | Observed                                                                                                                                                                                                                                                                                                                                                                       |
|--------------------------------|----------|----------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Apple App Store (apple.com/de) | local AI | 10 results                                         | 9 are local-AI apps by title: Locally AI (LM Studio), PocketPal AI, Liquid Apollo, Google AI Edge Gallery, Enclave, Oscilla, Atomic Chat, pocket, Draw Things. One (Grok) is not a local app. [S34]                                                                                                                                                                            |
| Google Play (Germany)          | local ai | 29 listings on the page as captured (16 PDF pages) | Titles such as LocalAI, Ollama Local AI, PocketPal AI, Off Grid AI, Private Mind, Gemma 4 Offline, Llamatik, Caelum, Secret AI. At least 4 listings name a connection to Ollama, vLLM, OpenRouter or LM Studio in their title or screenshot, so they may be clients for a server rather than on-device models. Several carry a rating of 2.2–4.2 stars; many carry none. [S35] |

**FACT** One documented search per store shows 10 results (Apple) and 29 listings (Google Play) for a local-AI query. **FACT** The author's statement "about ten iOS apps" matches the first Apple result page.

**FACT** For scale: the Apple App Store lists about 2.75 million apps and Google Play about 2.65 million (aggregator, 23 Oct 2026). [S36] Against that, a local-AI query surfaces

tens of apps. **ASSUMPTION** *Author:* it is very early in the cycle and there is basically nothing available yet for ordinary users.

**ASSUMPTION** **Made by:** Author (Hans Kuepper)

Much of the available local-AI software is distributed through GitHub, and GitHub is not a place for ordinary users, so for them it is effectively not available.

**Based on:** Test records (S22): GitHub is the vendor page, README or release source in five of the first eight tests (Page Assist, LTX Desktop, Unsloth, DiffusionBee, BoBe); LU's source was not checked. Page Assist's README leads with a source build that needs a terminal and Bun; the tester's first attempt failed on a PATH error (a Bun matter, not Page Assist itself) and he missed the store link lower on the page. GitHub topic pages list 5,794 public repositories for 'ollama' and 9,477 for 'local-llm' (S46); counts include libraries and forks, overlap between topics not removed. The author describes the field as thousands of forked projects where nobody knows what is current (author's observation, not measured). Limits: five test records, one tester.

**UNVERIFIED**

These are search result pages, not store-wide counts. The result number depends on the store's ranking and on how many results load on one page, so Apple 10 versus Google Play 29 is **not** a measure of which store has more local-AI apps. Titles are matched against the query, which inflates the number of listings named 'Local AI'. App function (on-device or server client, chat or other) was judged from titles and screenshots only; no app was installed.

**UNVERIFIED**

"Most iPhone apps are chatbot extensions." Not verified. Of the nine Apple results, one (Draw Things) is an image generator per the author's own test; the function of the others was not checked beyond their titles. The directory figure (74 of up to 315 listings are chat clients, all platforms) is the only count available.

**FACT**

Apple's WWDC 2026 announcements added image input to its Foundation Models framework and a Core AI framework for running other local models on Apple silicon.<sup>[S24]</sup> This provides on-device model access to third-party apps without the user installing a separate runtime.

## 4.4 Capital and corporate involvement

RUNTIME AND PLATFORM LAYER

| Company / segment | Reported funding or scale                                                                                                                                         | Source quality                                  |
|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------|
| Ollama            | \$65M Series B, 9 Jul 2026 (Theory Ventures; Benchmark, 8VC, Y Combinator and others); \$88M total; 14 employees; 8.9M monthly active developers (company-stated) | Multiple outlets; user number is company-stated |
| LM Studio         | ~\$19.3M over two rounds (May 2025)                                                                                                                               | Aggregator; unverified                          |
| Platform owners   | Apple (Foundation Models, Core AI), Microsoft (Copilot+ PCs), NVIDIA (DGX Spark)                                                                                  | Company announcements                           |

[S17, S18, S24]

APPLICATION LAYER (APPS FOR END USERS): WHAT THE SEARCH FOUND

| App / company                                                  | Funding trace found                                                                                                                                                                                                                  | Source quality                  |
|----------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------|
| GPT4All (Nomic AI)                                             | \$17M Series A, July 2023, led by Coatue (Y Combinator, Spark Capital and others). Nomic also sells embedding and data-mapping products, so the round is not purely for the app. <span>⚠ older than 1 year</span>                    | Aggregator page (startupintros) |
| AnythingLLM (Mintplex Labs)                                    | \$0.5M convertible note; investors Y Combinator, UpHonest Capital, Pioneer Fund. CB Insights lists the note at 1 Jan 2022, which pre-dates the product; date unverified. A second round is behind a paywall.                         | CB Insights, partial            |
| Jan, Msty Studio, Draw Things, Private LLM, Enchanted          | <b>No funding trace found.</b> Searches returned app-store pages, reviews and unrelated companies. Venture rounds are normally announced publicly; the author reads the absence as evidence of little investment (assumption below). | Web search only                 |
| DiffusionBee, PocketPal AI, BoBe, Page Assist, Caelum, Unsloth | Not searched.                                                                                                                                                                                                                        | -                               |

[S27, S28]

**Result.** **FACT** Funding is visible at the runtime layer (Ollama, LM Studio). At the consumer-application layer one older round (Nomic, 2023) and one small note (Mintplex) were found, and no trace for five other named apps; six more apps were not searched. [S17, S18, S27, S28] **AUTHOR-MEASURED** *Author:* one app he is in contact with has just found an investor; app and investor not named, not verified here.

**ASSUMPTION** **Made by:** Author (Hans Kuepper)

Very few investments are made in consumer local-AI applications. The market is early.

**Based on:** Venture investments are normally made public, and the search found none for five of eight named apps (LU, the ninth test, was not searched) and only one older and one small round for the other two. Limits: web search only; no paid database (Crunchbase, PitchBook) was queried; angel and unannounced rounds would not appear. The newly reported investor in one of the author's contact apps is one data point against a zero count and is not verified.

**FACT** A figure of "52 million monthly downloads" for Ollama circulates on secondary sites and conflicts with the company-stated 8.9 million monthly active developers; it is not used here.

## 4.5 Supply of applications: the vibe-coding argument

**ASSUMPTION** **Made by:** Author (Hans Kuepper)

Within one to two years, hundreds of thousands of apps will exist, including local-AI apps, because vibe coding lowers the cost of building software.

**Based on:** Author's direct experience building PromptQuorum applications with AI coding tools, and press reports about software companies being replaced by in-house builds. No time series of local-AI app counts was found to test the forecast.

**FACT** A figure of \$285 billion in software-stock value lost in 48 hours in early February 2026 is attributed to vibe-coding fears in one derivative article; the original could not be verified.<sup>[S25]</sup> It is not used as evidence in this report.

**FACT** Hugging Face reports that 1.5% of repositories account for 99.2% of downloads and that 85.6% of models have fewer than 200 lifetime downloads.<sup>[S19]</sup>

**ASSUMPTION**

*Inference from that distribution:* a large number of new apps does not by itself imply a large number of used apps; usage in open ecosystems is highly concentrated.

**Key takeaway.** The application layer is real but small and fragmented. Funding exists for runtimes, not clearly for consumer applications. Claims about the number of apps need a counted series, which the quarterly format can provide.

---

## 5. Local models versus frontier models

A frequent claim is that local models are six to twelve months behind the frontier, or only about four months for the largest. The author regards this as a marketing argument: it is offered as a reason to switch, and it does not describe what an ordinary machine can run. The source for the figures, Epoch AI, defines them for specific model classes.

**Epoch AI: lag of open-weight models behind closed frontier (months)**



Epoch Capabilities Index (ECI) for the first two rows; benchmark comparisons (GPQA-Diamond, MMLU-Pro, AA Intelligence, LMArena) for the consumer-GPU row, defined as up to 28B parameters (RTX 4090) or 40B (RTX 5090). The consumer-GPU analysis is older than 1 year. Epoch cautions that real-world usefulness lag is likely longer than benchmark lag because the best open model per benchmark is more likely to be over-fitted.

[S14, S15, S16]

- **FACT** Since January 2026 the most capable open-weight models lag the closed frontier by about four months; the earlier figure was three months. [S14, S16]
- **FACT** Models small enough for one consumer GPU reach frontier-level benchmark scores with a lag of 6–12 months. [S15]
- **FACT** Hugging Face reports that Chinese labs released the largest open model almost every month of 2026, with ceilings between 754 billion and 2.78 trillion parameters. [S19]
- **CALCULATION** A model of 754 billion parameters stored in 4-bit format needs more than 350 GB for weights alone (754 x 0.5 bytes), well above any consumer memory in Section 3. The 4-month figure therefore describes models few users can run.

**What the two lag figures actually cover**

| Model class         | Parameters                                                         | Memory for weights (4-bit, calculated) | Hardware needed                      | Measured lag behind frontier                              | Source   |
|---------------------|--------------------------------------------------------------------|----------------------------------------|--------------------------------------|-----------------------------------------------------------|----------|
| Largest open models | 754 billion to 2.78 trillion (Chinese labs' monthly largest, 2026) | about 377 GB to 1,390 GB               | Data centre or multi-machine cluster | about 4 months (Epoch Capabilities Index, since Jan 2026) | S14, S19 |

| Model class                                  | Parameters                             | Memory for weights (4-bit, calculated) | Hardware needed               | Measured lag behind frontier                                                    | Source |
|----------------------------------------------|----------------------------------------|----------------------------------------|-------------------------------|---------------------------------------------------------------------------------|--------|
| Models for one consumer GPU                  | up to 28B (RTX 4090) or 40B (RTX 5090) | about 14 GB / 20 GB                    | GPU with 24 GB / 32 GB memory | 6–12 months (benchmarks; <span style="color: red;">⚠ older than 1 year</span> ) | S15    |
| Models for a typical laptop or mid-range GPU | about 7–8B                             | about 4 GB                             | 8–16 GB total memory          | No figure found                                                                 | -      |

**CALCULATION** Weights only: parameters x 0.5 byte at 4-bit quantisation. Excludes context cache, runtime overhead and the operating system. Section 3.2 shows the system and open apps already use about 10–13 GB of a 24 GB Mac.

**In plain terms.** Two different lags are quoted and they describe different things. "About 4 months" compares the best open model with the best closed model; the best open models are too large for any consumer machine. "6–12 months" compares models small enough for *one consumer GPU with 24–32 GB of memory* with the frontier. Neither figure covers a computer with less memory.

**FACT** Epoch adds that the best-scoring open model on a benchmark is more likely to be over-fitted to it, so the lag in everyday usefulness is probably longer than the benchmark lag.<sup>[S15]</sup>

**ASSUMPTION** **Made by:** Author (Hans Kuepper)

The "local models are only months behind" statement is a marketing argument and cannot be used in practice by the reference user.

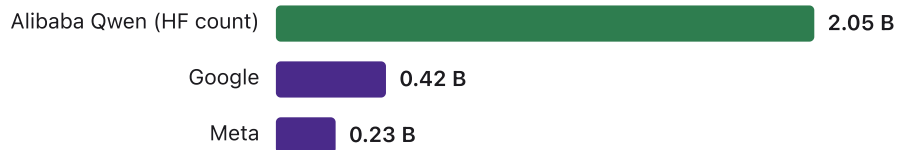
**Based on:** Table above: the 4-month figure covers models needing about 377 GB or more; the 6–12 month figure covers models that need a 24–32 GB GPU and is older than 1 year; no figure exists for a typical laptop or phone. Tests (S22): a 27B model on a 24 GB Mac ran out of memory on web-page reading and an event search; Caelum, a 1 GB model on a phone, gave a false news answer and a logic answer rated 5/25 by ChatGPT as judge (judge not independent, not verified). Limits: few tests, one tester, no side-by-side local-versus-cloud benchmark on the same tasks.

**Key takeaway.** The lag figures are real on benchmarks, but they describe models users cannot run (4 months) or older consumer-GPU data (6–12 months, older than 1 year). In the author's view they are a marketing argument, because for a typical laptop or phone no figure exists and the tests show failures.

## 6. Demand signals and what is changing

### 6.1 Developer-side adoption is rising

#### Hugging Face downloads in 2026, by model family (billions)



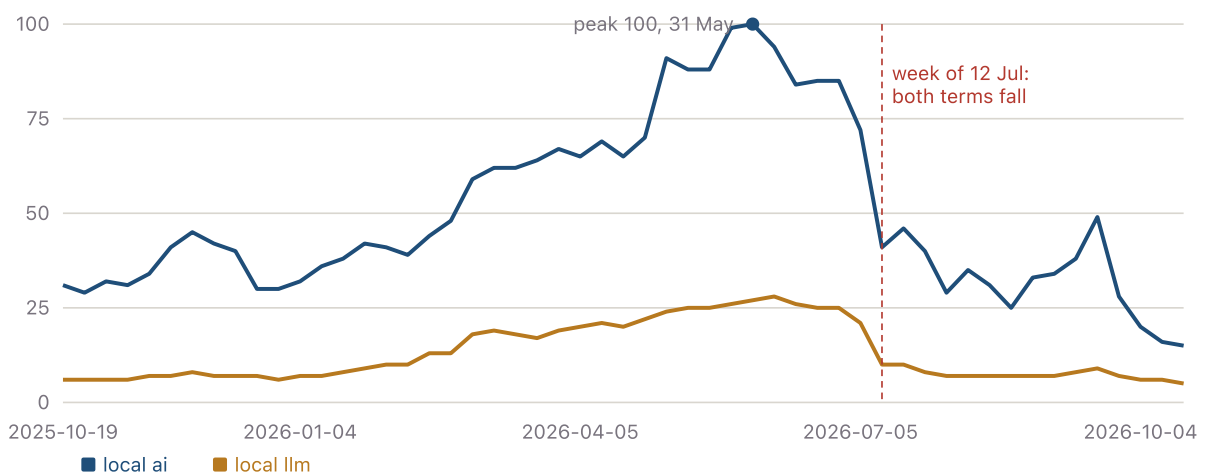
Sources: Hugging Face report as cited by Bloomberg and a French summary. Alibaba itself claims more than 3 billion in six months; Hugging Face's own count is lower. Periods are not identical (Google and Meta: 2026 to date; Qwen: Jan-Aug 2026). Download counts are not user counts.

[S19]

**FACT** Models under 1 billion parameters account for 83% of all-time Hugging Face downloads.<sup>[S19]</sup> **FACT** Ollama's developer base doubled between January and July 2026 and the company reports about one million new installs per week (company-stated).<sup>[S17]</sup>

### 6.2 Search interest

#### 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.

- **AUTHOR-MEASURED** **"local ai"**: about 31 in the first four weeks (Oct–Nov 2025); 100 in the week of 31 May 2026 (about 3.3 times the Oct 2025 level); 15 in the week of 18 Oct 2026, half the Oct 2025 level. Average of the last four weeks 19.8, which is 36% below the first four weeks (30.8).<sup>[S29]</sup>
- **AUTHOR-MEASURED** **"local llm"**: 6 in Oct 2025; peak 28 (week of 7 Jun 2026); 5 in the latest week. Roughly one quarter of the scale of "local ai" throughout.<sup>[S29]</sup>
- **AUTHOR-MEASURED** **The fall is abrupt, not gradual.** "local ai" went from 72 to 41 between the weeks of 5 Jul and 12 Jul (–43%); "local llm" from 21 to 10 (–52%) in the same week. Both had been flat at about 85 and 25 for the previous four weeks.<sup>[S29]</sup>
- **AUTHOR-MEASURED** From August to mid–September both terms sat at roughly their Oct 2025 levels (local ai 25–38), apart from one week at 49 (20 Sep), then fell again to 15.<sup>[S29]</sup>

#### UNVERIFIED

**The divergence cannot be explained.** Google Trends shows local-AI search interest falling from June and stepping down in July 2026, while other indicators rise: the Ollama developer base doubled between January and July (company–stated), Hugging Face downloads are high, and the author reports that other trends keep rising. The developer indicators available here end in July, so the August–October period cannot be compared. Both search terms moving by similar percentages in the same week fits a change in the data or in what is counted as well as a real change in interest; a search for a documented Google Trends methodology change in July 2026 found nothing. The export covers 12 months only, so a seasonal pattern cannot be excluded. No cause is claimed.

## 6.3 Privacy as a driver

**FACT** In a DuckDuckGo survey (n=1,944 US adults; the sponsor sells a privacy product), 32% of chatbot users had shared information they withheld from friends, relatives, colleagues or medical professionals; 53% did not know or were unsure that conversations may be used for AI training.<sup>[S20]</sup> **FACT** In a smaller academic survey (n=300), 63.6% reported privacy concerns about conversations with ChatGPT.<sup>[S21]</sup>

**ASSUMPTION** **Made by:** Author (Hans Kuepper)

Privacy is a weak sales argument for local AI because users have accepted data-for-service trades with Google and Meta for twenty years.

**Based on:** Author's observation of consumer behaviour. The surveys above show high stated concern together with high disclosure, which fits the author's view of behaviour, but they also show low awareness, which points to an information gap rather than to indifference. The data does not decide between the two.

## 6.4 Geography

UNVERIFIED

"China is a big local-AI market; other countries are underdeveloped; awareness is minuscule." No country-level data on local-AI users was found. What exists describes *model supply* (Chinese labs lead in the size of open releases and in Qwen downloads, and Hugging Face attributes 41% of platform downloads to Chinese models in the prior year) and does not measure where local AI is used.

### REVIEW OF THE "WAVE THEORY"

**FACT** The general concept is documented. Comin and Hobijn find that most technologies are introduced in advanced economies and then "trickle down" to lagging countries, and that diffusion sped up after World War II. <sup>[S31]</sup> **older than 1 year**

**FACT** For generative AI, adoption is uneven by country. Microsoft reports 17.8% of the world's working-age population using AI in Q1 2026, 27.5% in the Global North against 15.4% in the Global South, and a range from 70.1% in the UAE to 31.3% in the US (rank 21, up from 24). <sup>[S30]</sup>

| Part of the theory                                                    | Evidence                                                                                                                                  | Result     |
|-----------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|------------|
| Technology adoption is uneven across countries and reaches some later | Comin and Hobijn (S31); Microsoft 2026 country data (S30)                                                                                 | Confirmed  |
| New technology starts in advanced economies                           | Comin and Hobijn (S31); North-South gap in S30                                                                                            | Confirmed  |
| The US specifically leads each wave                                   | Supported for where AI models are built; <b>not</b> for usage: US ranks 21st in Microsoft's AI-use share, behind the UAE and others (S30) | Narrowed   |
| Local AI follows the same wave, and non-English markets lag           | No country-level data on local-AI use found; Google Trends (S29) is worldwide only                                                        | Unverified |

**ASSUMPTION** **Made by:** Author (Hans Kuepper)

Local AI spreads country by country in the same pattern as other technologies.

**Based on:** The general concept is backed by S30 and S31, so the theory is not "mad" and is kept. The step to local AI is an extrapolation without data. The "US leads" part is replaced by "advanced economies lead", because the AI-use ranking does not put the US first.

## 7. Trade and supply-chain implications (EU, China)

- **FACT** The memory shortage is global and driven by allocation to AI data-centre products; it affects EU buyers through higher PC and component prices. <sup>[S6, S7]</sup>
- **UNVERIFIED** A user post on NVIDIA's forum reports an EU marketplace price of 4,800 EUR for the 128 GB DGX Spark in February 2026 against \$4,699 in the US. Currency, VAT treatment and date were not verified. <sup>[S26]</sup>
- **FACT** The largest open-weight models and the most downloaded model family are published by Chinese labs; the Qwen family is described as Apache-2.0 licensed on one secondary source. <sup>[S19]</sup> For EU users this makes model *supply* dependent on Chinese labs, while the *hardware* supply chain (memory, GPUs) is dependent on US-designed and Asian-manufactured parts.
- **UNVERIFIED** EU regulatory effects (AI Act, data-protection rules) on local versus cloud deployment were not researched for this report.

## 8. Outlook, with assumptions labelled

| Statement                                                                                         | Type                 | Who / basis                                                            |
|---------------------------------------------------------------------------------------------------|----------------------|------------------------------------------------------------------------|
| Memory prices stay elevated into 2027                                                             | THIRD-PARTY FORECAST | IDC (Dec 2025). One secondary report cites a view to 2028; unverified. |
| AI PCs become the norm by 2029                                                                    | THIRD-PARTY FORECAST | Gartner (Aug 2025). NPU-based definition; pre-dates shortage.          |
| Future is hybrid; a growing share of AI moves local where privacy, cost or convenience justify it | ASSUMPTION           | Author. No measured local share exists; needs a tracked indicator.     |
| Local AI will grow faster in percent than the overall AI market                                   | ASSUMPTION           | Author. Plausible because it starts from a lower base; not tested.     |

| Statement                                                                   | Type       | Who / basis                                                                                |
|-----------------------------------------------------------------------------|------------|--------------------------------------------------------------------------------------------|
| Mobile will churn faster than PCs and Apple users will pay for on-device AI | ASSUMPTION | Author. Apple's WWDC 2026 changes are facts (4.3); the willingness to pay is not measured. |
| An on-device email organiser would succeed                                  | OPINION    | Author. No data; product opinion.                                                          |

## Projection: what could be ready in the coming years

| Driver           | What the sources say                                                                                                                                                                                                                               | Direction                            | Type                                            |
|------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------|-------------------------------------------------|
| Model efficiency | Capability density of open models doubled about every 3.5 months, Feb 2023 to Apr 2025 (51 models). A model with half the parameters matches the best open model about every 3.5 months. [S37]                                                     | Smaller models become more capable   | FACT for the period; extending it is ASSUMPTION |
| Memory supply    | Counterpoint: new output from 2027, more capacity in 2028, prices reversing by 2028. TrendForce: consumer prices keep rising next year. IDC: shortage into 2027. Samsung warning of worse in 2027–2028 (headline only, unverified). [S38, S7, S39] | Forecasts disagree                   | THIRD-PARTY FORECAST                            |
| AI PCs           | Gartner: 31% of shipments 2025, 54.7% 2026, 'the norm' by 2029. Definition is not about running large models; forecast pre-dates the memory shortage. [S8]                                                                                         | More devices with on-device AI chips | THIRD-PARTY FORECAST                            |
| Consumer GPUs    | RTX 60 series rumoured for H1 2027 or 2028; one leak gives up to 32 GB for the top card, equal to the 32 GB class of the RTX 5090 in S15. Unconfirmed. [S40, S15]                                                                                  | Unclear                              | UNVERIFIED                                      |
| App supply       | New app releases +60% (Q1 2026 versus Q1 2025) and +104% (April to date), Appfigures; +30% in 2025 to nearly 600,000, Sensor Tower. AI coding suspected, not measured. [S41, S42]                                                                  | More apps, quality unknown           | FACT for counts; cause ASSUMPTION               |

### ASSUMPTION Made by: Author (Hans Kuepper)

If local AI follows the general path of technology, capable hardware is likely to become available at mass-market prices over the coming years, apps are likely to get better, and AI coding tools are likely to speed up app development. Local AI is developing at great speed but is not there yet.

**Based on:** The table above: model efficiency doubling about every 3.5 months (measured to April 2025), memory supply expansion forecast from 2027–2028 by Counterpoint, Gartner's AI PC trajectory, and the rise in new app releases. Limits: memory forecasts conflict; the efficiency trend is a fit on 26 months of data and extending it to three or four years is an extrapolation; no source projects the share of

machines able to run a 30B model; the app data cover all apps, not local AI, and measure neither quality nor use.

## Proposed indicators to track every quarter

For the series to be quotable, each report should repeat the same measurements so that changes can be read as a time series.

| Indicator                                                         | Source                      | Q4 2026 value                                          |
|-------------------------------------------------------------------|-----------------------------|--------------------------------------------------------|
| Share of Steam systems with 16 GB+ VRAM; 24 GB+; 32 GB+           | Steam survey (primary page) | 16 GB: 25.9% (Jul); 24 GB: 5.43%; 32 GB: not retrieved |
| DGX Spark 128 GB list price; entry Apple Silicon price with 24 GB | NVIDIA, Apple               | \$6,950; to be recorded                                |
| Open-weight lag (months), frontier and consumer-GPU               | Epoch AI                    | ~4; 6–12 (older)                                       |
| Share of US households / consumers paying for AI                  | BofA, PNC, YipitData/a16z   | 2.2–4.5%                                               |
| Ollama monthly active developers                                  | Company                     | 8.9M (Jul 2026)                                        |
| Directory listings; share rated 3+ in hands-on tests              | PromptQuorum                | 315 upper bound; 3 of 7                                |

## 9. Claim register

Each claim in this report, with the result of the check.

| ID  | Claim                                                                            | Result               | Evidence                                                                                      | Source   |
|-----|----------------------------------------------------------------------------------|----------------------|-----------------------------------------------------------------------------------------------|----------|
| C01 | Few computers can run great local models; few have dedicated GPUs or 32 GB+ VRAM | Partly supported     | Steam Jul 2026: 51.4% have 10 GB+; 5.4% have 24 GB; 32 GB bucket not retrieved; gamer sample. | [S1]     |
| C02 | A 24 GB Mac uses ~10 GB with a few windows open                                  | Confirmed            | Measured 9.67 GB after startup, 12.92 GB with apps open.                                      | [S22]    |
| C03 | Even 32 GB VRAM may not suffice with other apps open                             | Opinion / assumption | Extrapolation; no 32 GB test.                                                                 | [S22]    |
| C04 | DGX Spark started at ~\$4,000, now \$6–7k                                        | Confirmed            | \$3,999 to \$4,699 (Feb) to \$6,950 (Oct, 128 GB); 64 GB at \$4,999; partner systems \$7–9k.  | [S3, S4] |

| ID  | Claim                                                                        | Result               | Evidence                                                                                                                                                                                                                    | Source                   |
|-----|------------------------------------------------------------------------------|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| C05 | High-end Apple computers cost like a small car did ten years ago             | Unverified           | Comparison not sourced. Top configurations cost several thousand dollars; no primary-source figure verified. Recommend dropping the comparison.                                                                             | -                        |
| C06 | 2–3% of US households pay for AI (credit card analysis)                      | Confirmed            | BofA ~3%, PNC 2.2%. Banks and card panels, not 'credit card companies'. al6z: 4.5% of consumers.                                                                                                                            | [S11, S12, S13]          |
| C07 | 97% of users use AI for free                                                 | Corrected            | Replaced by US card-data range: 2.2–4.5% of households or consumers pay. Menlo 2025 (~3%) is an implied revenue ratio; Menlo 2026 reports 55% of AI users pay (survey, different denominator, not used in the report body). | [S9, S10, S11, S12, S13] |
| C08 | Local AI is 'cheaper' but most people already use AI for free                | Opinion / assumption | Cost argument matters mainly to paying minority; top 1% = 19.5% of spend.                                                                                                                                                   | [S13]                    |
| C09 | Privacy is a weak argument; users accept Meta/Google trade-offs              | Opinion / assumption | Surveys show high stated concern, high disclosure, low awareness.                                                                                                                                                           | [S20, S21]               |
| C10 | Half the apps are underwhelming; the other half deliver but below frontier   | Partly supported     | Result of the author's own tests (linked in 4.2): 4 of 8 scored apps at 2 or below; n=9.                                                                                                                                    | [S22]                    |
| C11 | More than 75% of apps show no download time estimate                         | Corrected            | 60% of 5 evaluated apps (3 of 5); 3 of 8 at minimum.                                                                                                                                                                        | [S22]                    |
| C12 | No app tells the user the model's cut-off date                               | Unverified           | Not recorded in tests.                                                                                                                                                                                                      | [S22]                    |
| C13 | Apps list many models without size or purpose                                | Partly supported     | True for Draw Things, DiffusionBee; not for BoBe, Msty.                                                                                                                                                                     | [S22]                    |
| C14 | Local image apps cannot refine a single picture (cruise ship)                | Narrowed             | One app, one model; control misuse suspected; other edit tools untested.                                                                                                                                                    | [S22]                    |
| C15 | Image model could not render Heidelberg castle; ChatGPT (5) could a year ago | Partly supported     | First half observed in the DiffusionBee test. Second half is the author's recollection (ChatGPT 5); the earlier output was not kept.                                                                                        | [S22]                    |
| C16 | Local models are 6–12 months behind frontier                                 | Partly supported     | ~4 months for largest open models; 6–12 for consumer-GPU size; practical lag likely longer.                                                                                                                                 | [S14, S15]               |

| ID  | Claim                                                                   | Result               | Evidence                                                                                                                                                                                                                                                                                                                                                                     | Source               |
|-----|-------------------------------------------------------------------------|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| C17 | The 6–12 month claim assumes you have the right computer                | Confirmed            | Epoch limits it to models fitting 24–32 GB GPUs.                                                                                                                                                                                                                                                                                                                             | [S15]                |
| C18 | Google Trends interest in local AI fell substantially since summer 2026 | Confirmed            | 'local ai' peak 100 (31 May 2026), 15 on 18 Oct 2026; 36% below Oct–Nov 2025 on four-week averages; abrupt step in the week of 12 Jul. Unexplained: other indicators (Ollama developers, Hugging Face downloads, author's other trends) rise while Trends falls. Worldwide index, 12 months.                                                                                 | [S29]                |
| C19 | China is a big local-AI market; other countries minuscule               | Unverified           | Model supply data only; no user data.                                                                                                                                                                                                                                                                                                                                        | [S19]                |
| C20 | Technology trends spread in waves from the US                           | Narrowed             | Uneven diffusion from advanced economies is documented; the US is not first in AI usage (rank 21); local-AI pattern untested.                                                                                                                                                                                                                                                | [S30, S31]           |
| C21 | AI hardware gets more expensive, unlike normal tech                     | Confirmed            | Spark +74%; DRAM +49% QoQ; Dell/Lenovo +15–20%. Apple RAM upgrades unchanged in March.                                                                                                                                                                                                                                                                                       | [S3, S4, S5, S6, S7] |
| C22 | Hundreds of millions of capable devices by 2030                         | Partly supported     | Gartner AI-PC shipments 220.8M (2025–26 forecast); NPU-based, not LLM-capable.                                                                                                                                                                                                                                                                                               | [S8]                 |
| C23 | Mobile will churn faster; Apple users will pay for on-device AI         | Opinion / assumption | Not measured.                                                                                                                                                                                                                                                                                                                                                                | [S24]                |
| C24 | Hundreds of thousands of new apps within 1–2 years via vibe coding      | Opinion / assumption | No app-count series; HF downloads concentrated.                                                                                                                                                                                                                                                                                                                              | [S19]                |
| C25 | About ten local-LLM apps in the iOS App Store                           | Partly supported     | Apple DE search 'local AI': 10 results, 9 local-AI apps. A search page, not a store count.                                                                                                                                                                                                                                                                                   | [S34]                |
| C26 | Most iOS apps are chatbot extensions                                    | Unverified           | Not checked per app; 1 of 9 is an image generator. Directory: 74 of up to 315 listings are chat clients (all platforms).                                                                                                                                                                                                                                                     | [S23, S34]           |
| C27 | Little corporate or venture involvement                                 | Partly supported     | Runtime layer: funded (Ollama \$65M, LM Studio ~\$19M). Application layer: traces for 2 of 8 named apps, both small or old (GPT4All/Nomic \$17M, 2023; AnythingLLM \$0.5M note); none for 5. Author's inference: rounds are normally public, so few exist, early market. Author reports one contact app now has an investor (unverified, unnamed). No paid database queried. | [S17, S18, S27, S28] |

| ID  | Claim                                                                                                                                                   | Result               | Evidence                                                                                                                                                                                                                                | Source          |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| C28 | A local email-organising app would be a success                                                                                                         | Opinion / assumption | Product opinion.                                                                                                                                                                                                                        | -               |
| C29 | Local AI will grow faster than the AI market overall                                                                                                    | Opinion / assumption | Not tested.                                                                                                                                                                                                                             | -               |
| C30 | A local coding model could not read screenshots                                                                                                         | Narrowed             | True for Qwen3-Coder (text-only input); not for Qwen3-VL. Model used in the test to be confirmed.                                                                                                                                       | [S32, S33]      |
| C31 | Very few local-AI apps have a good, fully developed user interface                                                                                      | Partly supported     | In the author's 9 tests one app (LU) scored above 3 on onboarding (7 scored); four interface patterns recur. n=9, one tester, tester misreadings in at least 5 records; no data on typical users.                                       | [S22]           |
| C32 | Hardware and apps are likely to improve over the coming years if local AI follows the general technology path; AI coding tools speed up app development | Opinion / assumption | Supported directionally: efficiency trend (to Apr 2025), memory capacity forecast from 2027–2028 (forecasts conflict), new app releases +60–104%. No projection of machines able to run a 30B model; app data not specific to local AI. | [S37, S38, S41] |

## 11. Sources

Sources are listed with publication date. **⚠ older than 1 year** marks sources published before 23 October 2025. Secondary reports are named as such; where a primary page was not retrieved, this is stated.

- S1** Steam Hardware & Software Survey, July 2026 (secondary reports: tBreak, Hardware Busters, Club386). Valve primary page not retrieved. *Jul 2026*  
<https://tbreak.com/steam-hardware-survey-16gb-vram-overtakes-8gb/>
- S2** Steam survey, September 2026: platform share, macOS 1.92% of systems (Digital Citizen; TechSpot). Valve primary page not retrieved. *Sep 2026*  
<https://www.digitalcitizen.life/32gb-ram-overtakes-16gb-steam-hardware-survey-september-2026/>
- S3** NVIDIA developer forum: DGX Spark price change announcement (MSRP \$3,999 to \$4,699). Primary. *23 Feb 2026*  
<https://forums.developer.nvidia.com/t/2-23-2026-price-change-announcement/361713>
- S4** Implicator.ai: 64 GB DGX Spark at \$4,999; 128 GB at \$6,950; partner systems \$7,000–9,000 (NVIDIA primary not retrieved). *Oct 2026*  
<https://www.implicator.ai/nvidia-adds-64gb-dgx-spark-at-4-999-as-128gb-model-climbs-to-6-950/>

- S5** 9to5Mac: Apple M5 MacBook pricing, RAM upgrade prices unchanged. *3 Mar 2026*  
<https://9to5mac.com/2026/03/03/apple-macbook-price-increase-ram-same/>
- S6** Digital Citizen citing Sigmaintell: 16 Gb DRAM \$19.2 (Q1) to \$28.5 (Q2 2026). A second outlet (ProPakistani) states +89%; the underlying figures give +49%. *Jun 2026*  
<https://www.digitalcitizen.life/dram-and-ssd-prices-surge-again-as-memory-shortages-hit-pcs-and-phones/>
- S7** IDC via TechSpot: DRAM/NAND shortage into 2027; PC shipments could fall up to 8.9%. TrendForce via TweakTown: Dell/Lenovo +15–20%. *Dec 2025*  
<https://www.techspot.com/news/110679-memory-shortages-raise-pc-smartphone-prices-8-2026.html>
- S8** Gartner press release: AI PCs 31% of market 2025, 55% in 2026, 'the norm' by 2029. *28 Aug 2025* **⚠ older than 1 year**  
<https://www.gartner.com/en/newsroom/press-releases/2025-08-28-gartner-says-artificial-intelligence-pcs-will-represent-31-percent-of-worldwide-pc-market-by-the-end-of-2025>
- S9** Menlo Ventures, 2025 State of Consumer AI (\$12B spend, ~3% implied payers; survey of 5,031 US adults). *26 Jun 2025* **⚠ older than 1 year**  
<https://menlovc.com/perspective/2025-the-state-of-consumer-ai/>
- S10** Menlo Ventures, 2026 State of Consumer AI (55% of AI users report paying for at least one AI product). PDF text extract only. *Sep 2026*  
[https://menlovc.com/wp-content/uploads/2026/09/menlo\\_ventures\\_consumer\\_ai\\_report-2026.pdf](https://menlovc.com/wp-content/uploads/2026/09/menlo_ventures_consumer_ai_report-2026.pdf)
- S11** Bank of America Institute, 'Not quite mAlnstream' (~3% of BofA households pay for AI). *Mar 2026*  
<https://institute.bankofamerica.com/economic-insights/consumer-ai-usage.html>
- S12** TechDogs summarising PNC Consumer Report: 2.2% of PNC households paid for GenAI in May 2026 (secondary). *Jun 2026*  
<https://www.techdogs.com/tech-news/td-newsdesk/only-22-of-pnc-households-pay-for-genai-as-monthly-spend-climbs-to-31>
- S13** AI Weekly summarising a16z Top 100 Gen AI Consumer Apps (YipitData card panel): 4.5% of US consumers pay for ChatGPT/Gemini/Claude, top 1% of payers = 19.5% of spend (secondary). *Oct 2026*  
<https://aiweekly.co/alerts/a16z-top-1-of-us-ai-payers-drive-195-of-consumer-spend>
- S14** Epoch AI Data Insight: open-weight models lag closed frontier by ~4 months (ECI) since Jan 2026. *29 May 2026*  
<https://epoch.ai/data-insights/open-closed-eci-gap>
- S15** Epoch AI Data Insight: models runnable on a single consumer GPU match frontier with a 6–12 month lag (<=28B params for RTX 4090, <=40B for RTX 5090). Publication date not shown in retrieved copy; text refers to mid-2026 as a future date, so written in 2025. *2025 (undated)*  
**⚠ older than 1 year**  
<https://epoch.ai/data-insights/consumer-gpu-model-gap>

- S16** Epoch AI Data Insight: ~3 months average open-vs-closed gap, Jan 2023 – Oct 2025. *30 Oct 2025*  
<https://epoch.ai/data-insights/open-weights-vs-closed-weights-models>
- S17** SiliconANGLE / Pulse 2.0 / OpenSourceForU: Ollama \$65M Series B (Theory Ventures), \$88M total, 8.9M monthly active developers (company-stated). *9 Jul 2026*  
<https://siliconangle.com/?p=787680>
- S18** CB Insights: LM Studio ~\$19.3M raised (aggregator data, unverified by company). *2025* **older than 1 year**  
<https://www.cbinsights.com/company/lm-studio/financials>
- S19** Hugging Face Hub analysis Jan-Aug 2026 (via baeke.info); Qwen downloads (Business Standard/Bloomberg; EqualOcean). *14–18 Aug 2026*  
<https://news.baeke.info/articles/hugging-face-hub-analysis-chinese-labs-dominate-largest-open-models-as-small-models-drive-most-downloads>
- S20** DuckDuckGo survey, n=1,944 US adults, 17–27 Jun 2026 (vendor with commercial interest in privacy), via CyberInsider. *Jun-Jul 2026*  
<https://cyberinsider.com/duckduckgo-finds-one-third-of-ai-users-share-secrets-they-hide-from-people/>
- S21** arXiv 2508.06760: privacy norms around LLM chatbots, n=300 respondents. *Aug 2025* **older than 1 year**  
<https://arxiv.org/pdf/2508.06760>
- S22** PromptQuorum hands-on tests (author's own), nine tests published 9–24 Oct 2026. Index of test pages: <https://www.promptquorum.com/power-local-llm/>. Individual slugs: draw-things, bobe, page-assist, msty, unsloth, diffusionbee, ltx-desktop, caelum, locally-uncensored (suffix -hands-on-test). *9–21 Oct 2026*  
<https://www.promptquorum.com/power-local-llm/unsloth-hands-on-test>
- S23** PromptQuorum Local LLM Software Directory: category listing counts as of 23 Oct 2026 (author's own directory). *23 Oct 2026*  
<https://www.promptquorum.com/>
- S24** Callstack: on-device AI after WWDC 2026 (Foundation Models with image input, Core AI). *2026*  
<https://callstack.com/blog/on-device-ai-after-wwdc-2026-whats-new>
- S25** TechTarget-derived article (via kapernikov mirror): '\$285 billion' software-stock loss in early Feb 2026 attributed to vibe-coding fears. Weak provenance; not independently verified. *2026*  
<https://lidarmonitor.kapernikov.com/news/vibe-coding-what-it-leaders-need-to-know-techtarget>
- S26** NVIDIA developer forum thread: user-reported EU marketplace price of DGX Spark 4,800 EUR (unverified). *Feb 2026*  
<https://forums.developer.nvidia.com/t/dgx-spark-price-increase/361640>
- S27** Startupintros: Nomic AI (GPT4All) \$17M Series A led by Coatue (aggregator page; primary announcement not retrieved). *1 Jul 2023* **older than 1 year**  
<https://startupintros.com/news/2023-07-01-nomic-ai-series-a>

- S28** CB Insights: Mintplex Labs (AnythingLLM) \$0.5M convertible note; investors Y Combinator, UpHonest Capital, Pioneer Fund (partial, second round paywalled). *2022 (round date as listed)*  
**⚠ older than 1 year**  
<https://www.cbinsights.com/company/mintplex-labs/financials>
- S29** Google Trends export supplied by the author: terms 'local llm' and 'local ai', Worldwide, All categories, weekly, 19 Oct 2025 to 18 Oct 2026 (53 weeks). Index 0–100 relative to the highest point of the two terms. *export 23 Oct 2026*  
<https://trends.google.com/trends/explore>
- S30** Microsoft, 'The state of global AI diffusion in 2026': 17.8% of the world's working-age population used AI in Q1 2026 (16.3% before); Global North 27.5%, Global South 15.4%; UAE 70.1%; US 31.3%, rank 21 (was 24). Method not reviewed here. *7 May 2026*  
<https://blogs.microsoft.com/on-the-issues/2026/05/07/the-state-of-global-ai-diffusion-in-2026/>
- S31** Comin and Hobijn, New York Fed Staff Report 169 (cross-country technology adoption): new technologies typically appear in advanced economies and 'trickle down' to lagging countries; diffusion sped up after World War II. Publication year not retrieved; the report series number places it well before 2026 (assumption). *year not retrieved* **⚠ older than 1 year**  
[https://www.newyorkfed.org/medialibrary/media/research/staff\\_reports/sr169.html](https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr169.html)
- S32** Ollama model library, qwen3-vl: described as the most powerful Qwen vision-language model; accepts image and video input. *captured 23 Oct 2026 (page shows '5 months ago')*  
<https://registry.ollama.ai/library/qwen3-vl:2b>
- S33** Ollama model library, qwen3-coder: listed input type is text only; sizes 30B (19 GB) and 480B (290 GB). *captured 23 Oct 2026*  
<https://ollama.com/library/qwen3-coder>
- S34** Apple App Store search 'local AI' (apple.com/de), screenshot PDF supplied by the author, captured 23 Oct 2026 15:23: 10 results. *23 Oct 2026*  
<https://www.apple.com/de/search/>
- S35** Google Play search 'local ai' (Germany, German interface), screenshot PDF supplied by the author, captured 23 Oct 2026 15:22: 16 pages, 29 app listings. *23 Oct 2026*  
<https://play.google.com/store/search?q=local%20ai&c=apps&hl=de>
- S36** 42matters store statistics: 2,749,633 apps in the Apple App Store and 2,651,095 in Google Play (page dated 23 Oct 2026, aggregator). *23 Oct 2026*  
<https://data.42matters.com/stats>
- S37** Tsinghua University: Densifying Law of LLMs (Nature Machine Intelligence cover article). Maximum capability density of open-source LLMs roughly doubles every 3.5 months; a model with half the parameters can match the state of the art about every 3.5 months. Dataset: 51 open models, Feb 2023 to Apr 2025. Assumes models trained equally well have the same density; describes the best open models, not typical ones. *issue of 20 Nov (year not stated on page); data to Apr 2025*  
**⚠ older than 1 year**  
<https://www.tsinghua.edu.cn/en/info/1245/14611.htm>

- S38** BigGo Finance report on a Counterpoint Research forecast (Kang Kyung-soo, iForum 2026, 22 Oct 2026): new memory output from SK Hynix and Micron from 2027, Samsung and CXMT expansions ramp in 2028; headline cites prices reversing to a 20–30% decline by 2028. Same article: TrendForce expects consumer memory prices to keep rising next year; Q3 2026 DRAM contract prices +13–18% quarter on quarter. *22 Oct 2026*  
<https://finance.biggo.com/news/c3a3fef8-8dfc-4a8b-9acf-c3ae851ea8c8>
- S39** TweakTown headline: 'Memory shortages will get only worse in 2027 and 2028, warns Samsung'. Article could not be retrieved; headline only, treat as unverified. *2026 (date not retrieved)*  
<https://www.tweaktown.com/news/112966/memory-shortages-will-get-only-worse-in-2027-and-2028-warns-samsung/index.html>
- S40** God Is A Geek: RTX 6090 rumoured for H1 2027 (Moore's Law Is Dead, 14 Sep 2026); leakers dispute it and suggest 2028; up to 32 GB GDDR7 in unattributed leaks. NVIDIA has announced nothing. *Sep 2026*  
<https://godisageek.com/2026/09/nvidia-rtx-6090-h1-2027-rumor/>
- S41** TechCrunch, citing Appfigures: new app releases Q1 2026 versus Q1 2025 +60% worldwide (+80% on iOS); April 2026 to date +104% (+89% iOS). AI coding tools are the suspected cause, not measured. *18 Apr 2026*  
<https://techcrunch.com/2026/04/18/the-app-store-is-booming-again-and-ai-may-be-why/>
- S42** 9to5Mac, citing The Information and Sensor Tower: new App Store apps grew 30% to nearly 600,000 in 2025 versus 2024; the report suggests vibe-coding tools are the main contributors but cannot say how many apps were AI-built. *6 Apr 2026*  
<https://9to5mac.com/2026-04-06/app-store-sees-84-surge-in-new-apps-as-ai-coding-tools-take-off/>
- S43** IDC and Omdia Q3 2026 PC shipments, as reported by Digital Today (Korea) and Windows Latest: IDC -20.1% to 62.7M units (from 78.5M); Omdia -21.2% to 58.1M (laptops, desktops, workstations). Both attribute the fall to the AI-driven memory squeeze. Neither article gives an AI-PC share. *11 Oct 2026*  
<https://www.windowslatest.com/2026/10/11/ai-was-supposed-to-sell-pcs-instead-pc-shipments-just-crashed-20-as-ai-eats-the-worlds-ram-and-windows-11s-copilot-agents-arent-convincing-buyers/>
- S44** Canalsys, as reported by Vivre and Computer Weekly: AI-capable PCs were 14% of global shipments in Q2 2024 (8.8M units) and 20% in Q3 2024 (13.3M units). Definition: a desktop or notebook with a chipset or block for dedicated AI workloads, such as a neural processing unit (NPU). *Aug and Nov 2024* **older than 1 year**  
<https://www.computerweekly.com/news/366615506/A-fifth-of-new-PCs-shipped-this-quarter-are-AI-optimised>
- S45** Tom's Hardware, 'What is an AI PC': Microsoft Copilot+ PC requires an NPU of at least 40 TOPS, 16 GB RAM and 256 GB storage; Intel defines an AI PC as having a CPU, GPU and NPU with AI acceleration. The article does not say whether these PCs can run large language models locally. *15 Jun 2024* **older than 1 year**  
<https://www.tomshardware.com/laptops/what-is-an-ai-pc>

**S46** GitHub topic pages, public repositories matching the topic: 'ollama' 5,794 and 'local-llm' 9,477 (page counts, 25 Oct 2026). Counts include libraries, tools and forks, not only end-user apps; overlap between topics not removed. 25 Oct 2026  
<https://github.com/topics/local-llm>

**Key takeaway.** Capable hardware is scarce and getting costlier; applications are many, uneven and hard to compare; the capability gap to frontier is small for giant models and larger for models a normal machine can run. All three are measurable each quarter, which is what the series should track.

Free to quote and reuse under [CC BY 4.0](#); credit Hans Kuepper and PromptQuorum, link to <https://www.promptquorum.com/about/state-of-local-ai/2026-q4> and say if you changed anything. Third-party sources stay under their owners' terms.