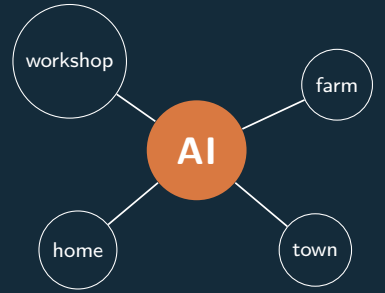
AI After Peak Energy

**When energy tightens,
who owns useful intelligence?**

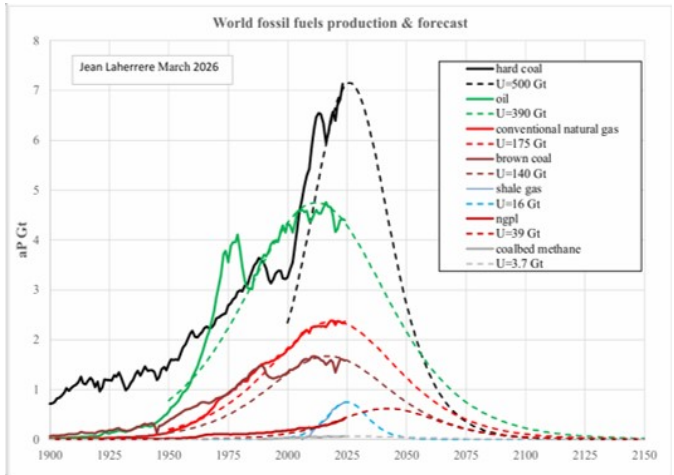


Peak Oil discussion / Andrii Zvorygin

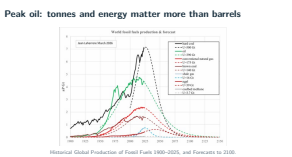
Energy, computing, and local intelligence

A kitchen table has a meeting recording, a repair question, or a crop plan. A clerk needs the bylaw, a mechanic needs the manual, and a grower needs weather and soil records. Each task depends on electricity, chips, buildings, cooling, and people. As energy tightens, the useful question is ownership: who keeps the tools, records, and skill needed to continue the work? This story moves from physical energy to computing and back to the local table, where a machine matters when it helps someone repair, decide, teach, or grow.

Peak oil: tonnes and energy matter more than barrels



Historical Global Production of Fossil Fuels 1900–2025, and Forecasts to 2100.



The chart shows oil, gas, coal, and the other large fuel streams already past their estimated production peaks; only the smallest NGPL stream peaks later. The forecasts then carry those streams downward across the century. Pressure reaches transport, food, heat, and public works through the energy system itself. A town still needs those services every day. The practical question is how to keep useful work moving when each unit of energy carries more weight.

When energy gets tighter, every sector seeks efficiency

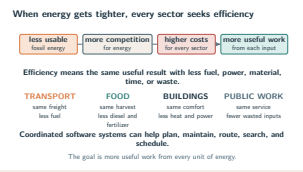
less usable fossil energy → **more competition** for energy → **higher costs** for every sector → **more useful work** from each input

Efficiency means the same useful result with less fuel, power, material, time, or waste.

TRANSPORT	FOOD	BUILDINGS	PUBLIC WORK
same freight less fuel	same harvest less diesel and fertilizer	same comfort less heat and power	same service fewer wasted inputs

Coordinated software systems can help plan, maintain, route, search, and schedule.

The goal is more useful work from every unit of energy.



Picture a municipal crew planning a route, a grower planning a harvest, and a clerk scheduling maintenance. Each person needs the same useful result with less fuel, power, material, time, and waste. Coordinated software can search records, compare options, route work, and schedule the next step. People still choose the plan and do the physical work. The gain is more useful work from each scarce input, which leads naturally to the co-operation communities use to keep essential services moving.

Pressure creates new levels of co-operation

CLIMATE + WATER + ENERGY PRESSURE → SHARED WORK → NEW INSTITUTIONS

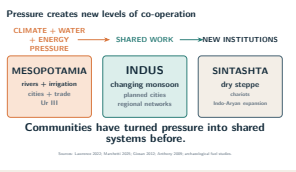
MESOPOTAMIA
rivers + irrigation
cities + trade
Ur III

INDUS
changing monsoon
planned cities
regional networks

SINTASHTA
dry steppe
chariots
Indo-Aryan expansion

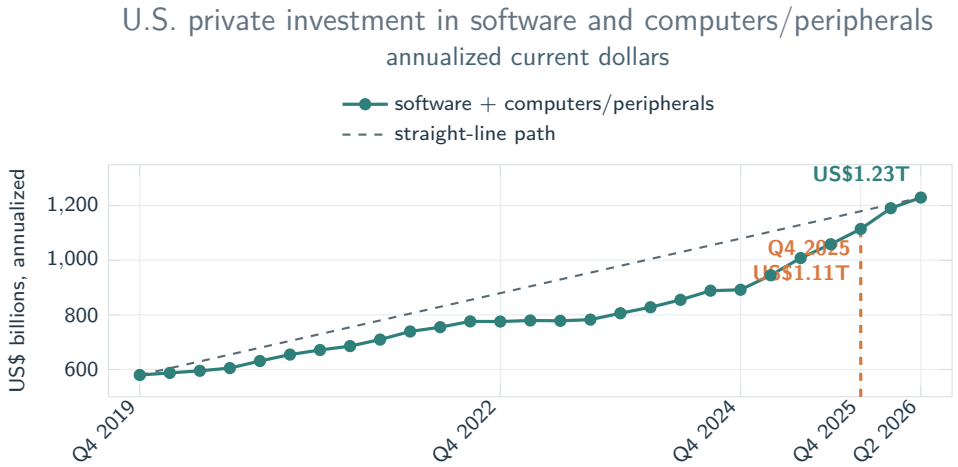
Communities have turned pressure into shared systems before.

Sources: Lawrence 2022; Marchetti 2025; Giosan 2012; Anthony 2009; archaeological fuel studies.

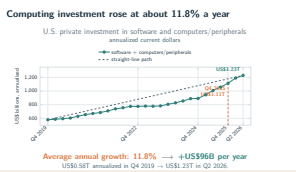


Three communities faced material pressure in a world powered by sunlight, food, wood, and animal labour. Climate shifts changed water and biomass; declining tree cover and scarcer fuel reduced available energy. Mesopotamian towns organized rivers, irrigation, and fuel; Indus settlements coordinated around changing water systems; Sintashta communities joined metalworking, horses, chariots, and long-distance exchange during Indo-Aryan expansion. Shared work became infrastructure in each case. The later question is whether computing can become another shared capability close to the work.

Computing investment rose at about 11.8% a year

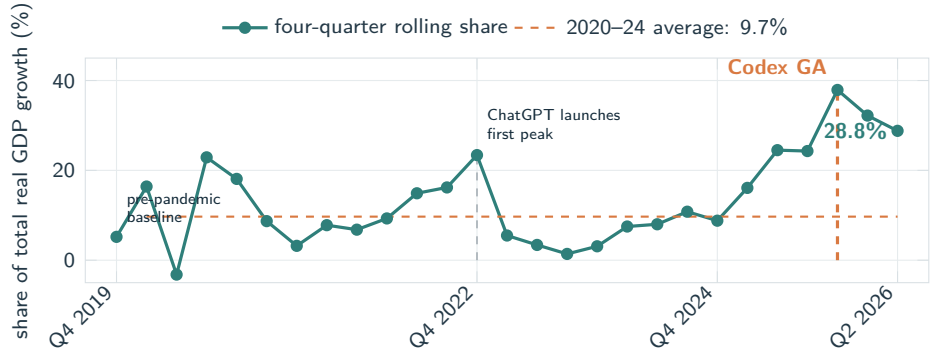


Average annual growth: 11.8% ⟶ +US$96B per year



This first chart sets the scale. Private investment in software plus computers and peripherals rises in a fairly straight upward path from about US$580 billion annualized in Q4 2019 to US$1.23 trillion in Q2 2026. The average annual growth rate over that period is about 11.8%; the straight-line path rises by about US$96 billion per year. The series shows the capital entering computing before we ask how large that investment was relative to the rest of the economy.

Share dips when the rest of GDP catches up



37.9% in Q4 2025 ⟶ 28.8% in Q2 2026

Pandemic computerization raised the installed base; later dips show broader output catching up.

BEA software and computers/peripherals contributions via FRED; latest revised data through Q2 2026.

The pandemic caused a burst of computer purchases for remote work and digital services. That equipment stayed in use, so computing investment kept rising. Once the new computers were in place, the rest of the economy began growing faster, and computing made up a smaller share of total growth. The same pattern followed ChatGPT and the Q4 2025 coding-agent release: computing's share reached 37.9 percent, then fell to 28.8 percent in Q2 2026. The first chart shows computing spending rising; this chart shows the rest of the economy catching up.

What people call AI: modern software

A trained model is a highly compressed database of patterns; tools retrieve current evidence and run tests.

MODEL MEMORY

a compressed database of patterns
language · code · images
video · audio
learned relationships in the weights

GENERATION

language models: next-token sequence
image/video models: repeated denoising
vision, speech, and music use their own model forms

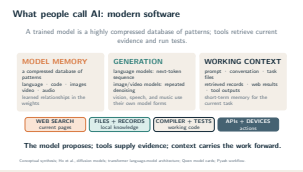
WORKING CONTEXT

prompt · conversation · task files
retrieved records · web results · tool outputs
short-term memory for the current task

WEB SEARCH current pages | **FILES + RECORDS** local knowledge | **COMPILER + TESTS** working code | **APIs + DEVICES** actions

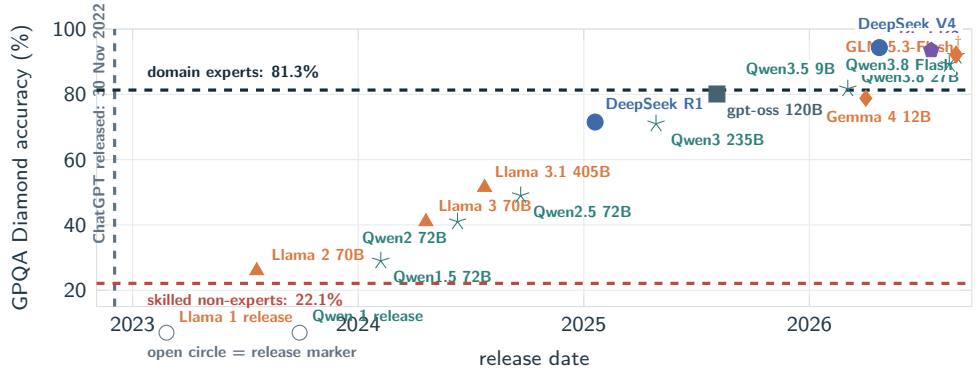
The model proposes; tools supply evidence; context carries the work forward.

Conceptual synthesis; Ho et al., diffusion models; transformer language-model architecture; Qwen model cards; Pyash workflow.

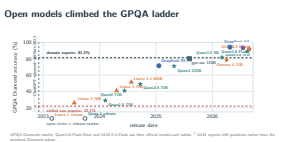


Modern software has several parts. A model is a compressed database of patterns learned from text, code, images, video, speech, and other records; its weights store relationships used to generate outputs. Different models use different methods: language models predict tokens, diffusion models denoise images or video, and vision, speech, music, and coding systems have their own forms. Tools connect a model to current web pages, local files and databases, compilers, and tests. The task's prompt, retrieved records, and tool results form short-term working memory; saved records, tests, and later training preserve institutional knowledge.

Open models climbed the GPQA ladder

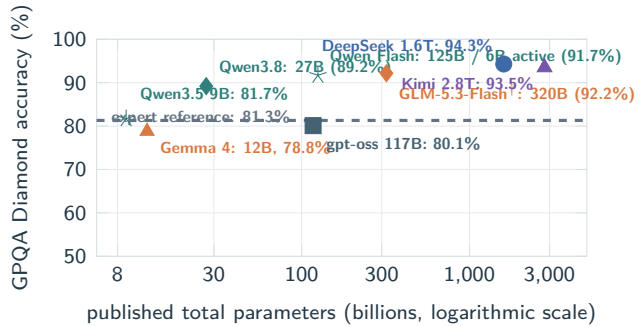


GPQA Diamond results; Qwen3.8-Flash-Next and GLM-5.3-Flash use their official model-card tables; † GLM reports 448 questions rather than the standard Diamond subset.



The open-model points climb the ladder. GPQA asks difficult science and mathematics questions and compares answers with people who have graduate-level subject training. Qwen3.5-9B reaches the human reference line, while Qwen3.8 Flash-Next reaches 91.7% on its model-card result. That changes the local decision. A downloadable model can now handle broad knowledge work on owned equipment, while tools, records, tests, and accountable people shape the result for a town.

The last few points require much larger models



GPQA model-card results; † GLM-5.3-Flash uses the separate 448-question table. Total parameters are used for the horizontal scale.

Qwen3.5: 9B 81.7% ≈ GPQA expert

Qwen3.8: 27B 89.2% GPQA

Qwen3.8 Flash-Next 125B total / 6B active / 91.7% GPQA

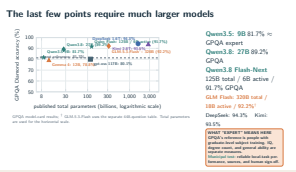
GLM Flash: 320B total / 18B active / 92.2%†

DeepSeek: 94.3% Kimi: 93.5%

WHAT "EXPERT" MEANS HERE

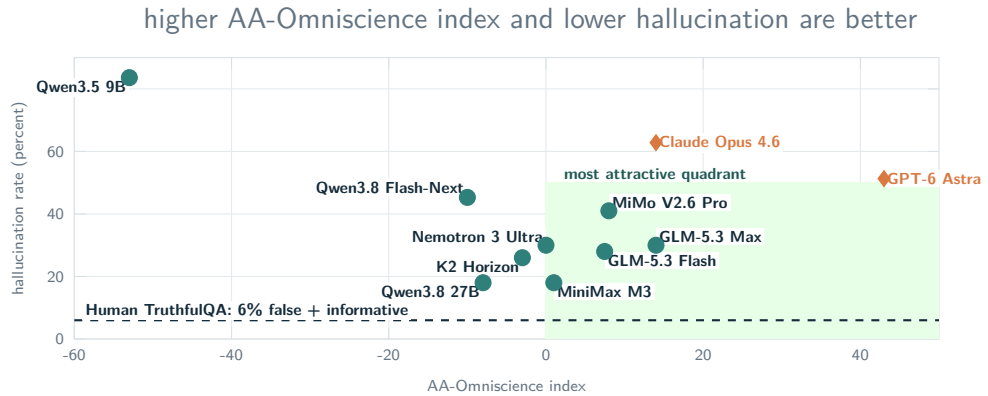
GPQA's reference is people with graduate-level subject training. IQ, degree count, and general ability are separate measures.

Municipal test: reliable local-task per-



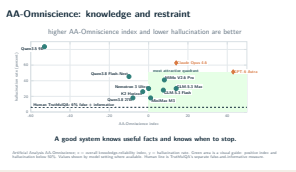
At the top of the ladder, each extra point requires a much larger system. Here, "expert" means people with graduate-level training in the tested subject. A municipality needs a more useful local standard: reliable answers from its own records, tested against known examples, with an accountable person reviewing the result. That turns a broad benchmark into a decision about one real task, such as a bylaw search, a repair instruction, or a public notice.

AA-Omniscience: knowledge and restraint



A good system knows useful facts and knows when to stop.

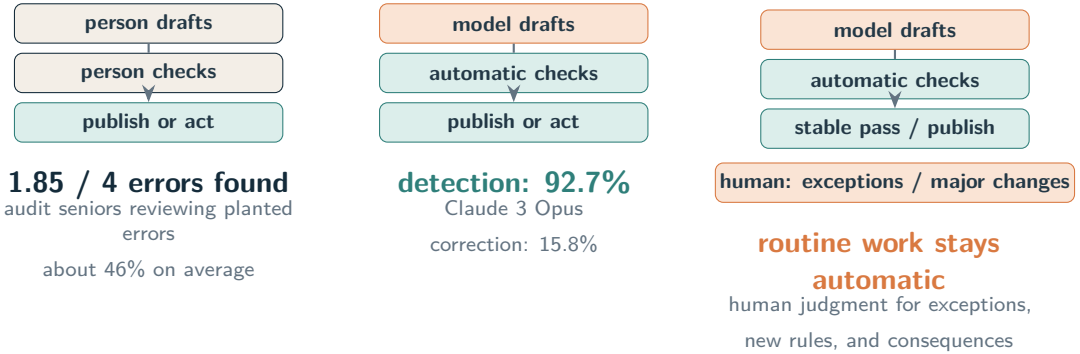
Artificial Analysis AA-Omniscience; x = overall knowledge-reliability index, y = hallucination rate. Green area is a visual guide: positive index and hallucination below 50%. Values shown by model setting where available. Human line is TruthfulQA's separate false-and-informative measure.



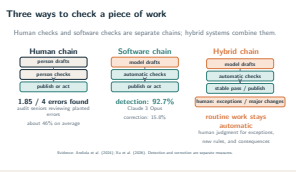
AA-Omniscience combines factual recall with restraint: correct answers raise the index, false guesses lower it, and refusing an unknown question carries no penalty. The vertical axis isolates the false-guess rate. Qwen3.8 27B sits at minus eight and 18 percent without reasoning; Flash-Next sits at minus ten and about 45 percent. GLM-5.3 Flash reaches a positive 7.47 index with 28 percent hallucination, placing it in the attractive quadrant with several newer open models. GLM-5.3 Max, Nemotron 3 Ultra, K2 Horizon, MiniMax M3, and MiMo V2.6 Pro provide additional open-model reference points. The hosted points show the same trade-off at much higher index values. The human line comes from TruthfulQA's separate six-percent false-and-informative rate.

Three ways to check a piece of work

Human checks and software checks are separate chains; hybrid systems combine them.

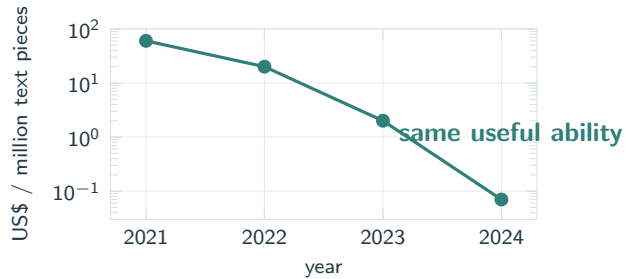


Evidence: Andiola et al. (2024); Xu et al. (2026). Detection and correction are separate measures.



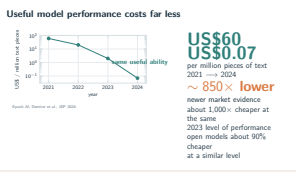
A clerk prepares a public notice. In a human chain, one person drafts and another checks; the cited audit found people caught about 46% of planted errors. In a software chain, the system drafts and automatic checks run every time; the cited model study detected 92.7% of planted errors and corrected 15.8%. A hybrid chain keeps routine passes automatic and brings a person to exceptions, new rules, major changes, and consequences.

Useful model performance costs far less



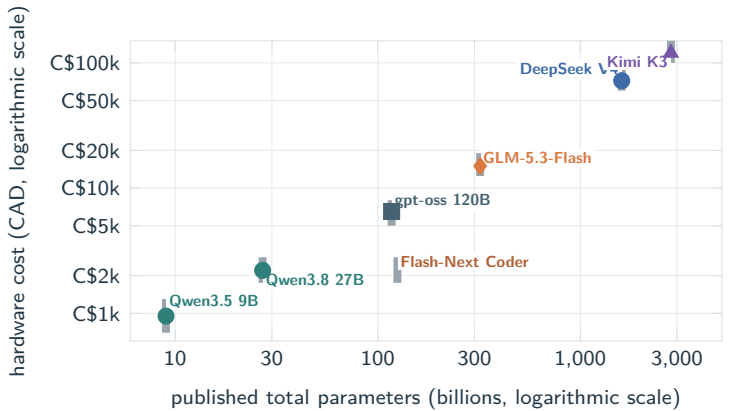
Epoch AI; Demirer et al., JEP 2026.

US$60
US$0.07
per million pieces of text
2021 ⟶ 2024
∼ 850× lower
newer market evidence
about 1,000× cheaper at
the same
2023 level of performance
open models about 90%
cheaper
at a similar level



The cost axis descends from the benchmark line. Smaller models, quantisation, serving software, and better hardware bring useful performance within reach of smaller owners. The full price includes servers, power, buildings, staff, spare capacity, research, and profit. A municipality can choose a bounded job, price the complete system, and buy only the capability that job requires. Useful local work can become affordable while frontier-scale systems remain expensive.

The local bill is mostly hardware



4-bit weights + 20% memory headroom; used RTX 3060-class 12GB cards at about C$450, RTX 3090-class 24GB cards at about C$1,100. Hardware purchase only.

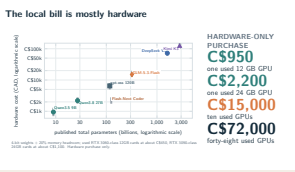
HARDWARE-ONLY PURCHASE

C$950
one used 12 GB GPU

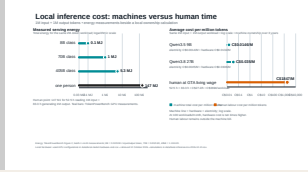
C$2,200
one used 24 GB GPU

C$15,000
ten used GPUs

C$72,000
forty-eight used GPUs



The equipment list makes the ownership difference clear. Small Qwen systems and the compressed Flash-Next Coder fit the one-card ownership path; larger models need several coordinated cards. A mixture-of-experts model activates part of its weights per token, while the full model still sets the memory requirement. The first local bill is the computer itself. Power, cooling, networking, maintenance, and staff turn that computer into a dependable public service.

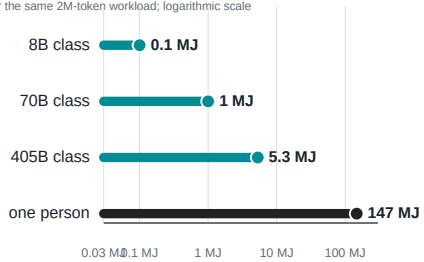


Local inference cost: machines versus human time

1M input + 1M output tokens • energy measurements beside a local ownership calculation

Measured serving energy

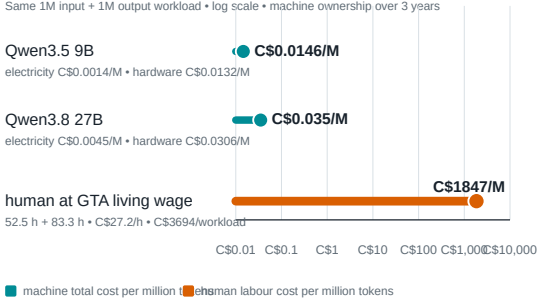
Total energy for the same 2M-token workload; logarithmic scale



Human point: 147 MJ for 52.5 h reading 1M input +
83.3 h generating 1M output. Teal bars: TokenPowerBench GPU measurements.

Average cost per million tokens

Same 1M input + 1M output workload • log scale • machine ownership over 3 years



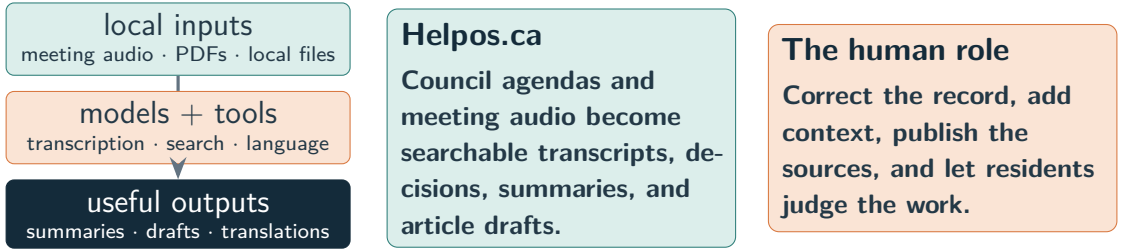
Machine line = hardware + electricity; log scale.
At 100 workloads/month, hardware cost is ten times higher.
Human labour remains outside the machine bill.

Energy: TokenPowerBench Figure 2, batch-1 vLLM measurements; 8B = 0.02/0.08 J input/output token, 70B = 0.20/0.80, 405B = 1.10/4.20.

Local hardware: used-GPU configurations in data/local-model-hardware-cost.csv • retrieved 10 October 2026 • calculations in data/local-inference-tco-2026-10-10.csv.

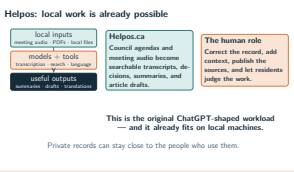
The same workload reaches a municipal desk. A person must read the source, produce the text, check the result, and repeat that work across the records. The hours add up quickly. The chart places that labour beside machine energy and ownership cost, with hardware, power, maintenance, and human checking kept visible. Helpos gives us a real local example: the value comes from many small records becoming searchable, clear, and useful to residents.

Helpos: local work is already possible



This is the original ChatGPT-shaped workload — and it already fits on local machines.

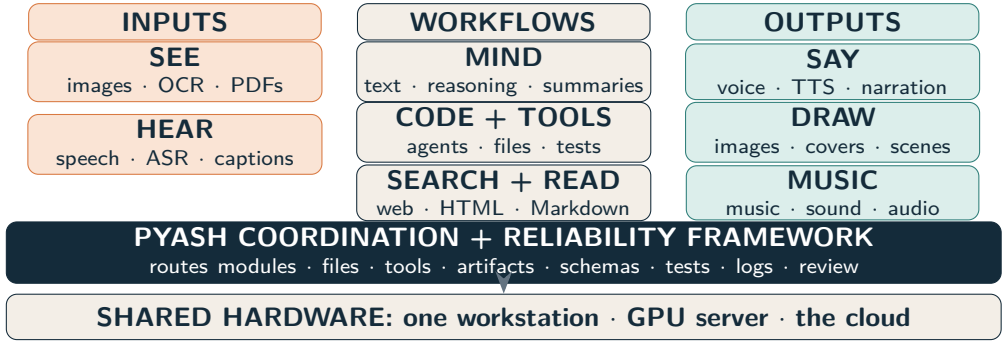
Private records can stay close to the people who use them.



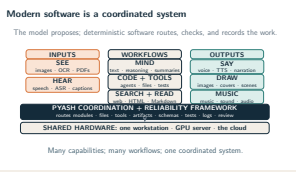
A council recording reaches a municipal desk. A local speech recognizer produces the transcript, a language model searches a bylaw and summarizes the discussion, and the software prepares a public-information draft. Helpos.ca shows this workload running close to the people who use it. Staff correct the record, approve publication, and keep responsibility for the decision. Residents can inspect the sources, follow the record, and judge the result.

Modern software is a coordinated system

The model proposes; deterministic software routes, checks, and records the work.



Many capabilities; many workflows; one coordinated system.



Helpos contains separate jobs. Vision models read images and OCR, speech models hear recordings, language models search and summarize, coding agents edit files, and image, voice, and music models create outputs. Model output is the soft part: it proposes text, code, or a decision. Pyash supplies the firmer frame around it through routing, schemas, rules, tests, logs, permissions, artifacts, and review. Together, those parts form a coordinated software system. One computer can host many capabilities when the workflow routes each job to the right model and checks the result.

What the local work would cost in human time

Combined Helpos/local-work audit: recorded work completed on owned machines

Helpos archive: 174 transcript records · 75 agenda pages
Audit: 3,807 Ollama responses · 203 transcript outputs · 2,149 images

TEXT WORK

9.00M input + 1.32M output tokens

582 hours

C$15,829 human-labour equivalent

TRANSCRIPTION

308.84 audio-hours

1,235 hours

C$33,602 at 4 labour-hours per audio hour

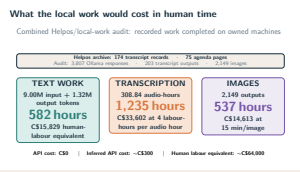
IMAGES

2,149 outputs

537 hours

C$14,613 at 15 min/image

API cost: C$0 | **Inferred API cost: ∼C$300** | **Human labour equivalent: ∼C$64,000**



The Helpos archive contains 174 transcript records, 75 agenda pages, 3,807 local text responses, 203 transcript outputs, and 2,149 images. At the stated labour assumptions, that is about 2,354 hours, or roughly C$64,000 of human time. Owned machines paid no API fees, while hardware and electricity remain part of ownership. This gives the local-service baseline: the town already has a substantial body of work to measure before we turn to coding.

Workflow helps — this study used GPT-4o

A workflow result using a model released in May 2024.

EARLY 2025 · MATURE PROJECTS

METR randomized study

16 experienced developers · 246 tasks

19% longer

Cursor + Claude 3.5/3.7
reviewing and cleaning
output absorbed time

2026 · WORKFLOW STUDY

CHI controlled study

GPT-4o backend
released May 2024

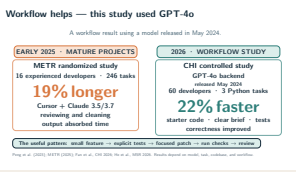
60 developers · 3 Python tasks

22% faster

starter code · clear brief · tests
correctness improved

The useful pattern: small feature → explicit tests → focused patch → run checks → review

Peng et al. (2023); METR (2025); Fan et al., CHI 2026; He et al., MSR 2026. Results depend on model, task, codebase, and workflow.



An experienced developer opens a mature repository with an early coding assistant. METR's study found that this group took longer on its assigned tasks, while a later controlled study found faster work on three small Python tasks with a GPT-4o backend released in May 2024. The famous Copilot result measured one contained JavaScript task. These results point to a workflow requirement: define the behaviour, write the tests, keep the change focused, and review the result.

Q4 2025 changed what stays cloud-heavy

Chat-era work

Prompt ⟶ answer
transcribe audio
summarize a file
draft an article
translate or explain
search a local archive

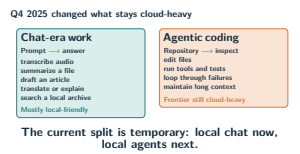
Mostly local-friendly

Agentic coding

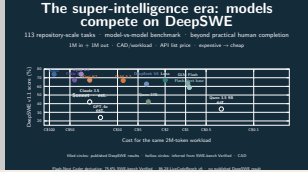
Repository ⟶ inspect
edit files
run tools and tests
loop through failures
maintain long context

Frontier still cloud-heavy

The current split is temporary: local chat now, local agents next.



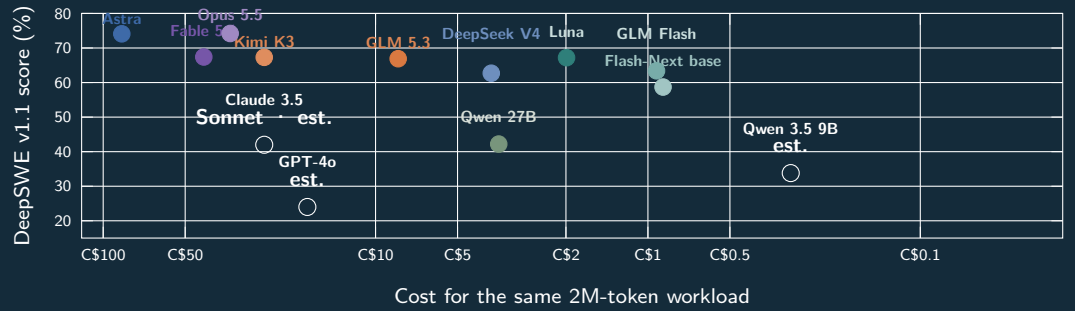
A local clerk can already give a model a recording, archive, report, or article and receive useful text. By Q4 2025, coding agents could inspect a repository, edit files, run tools and tests, and work through failures. That expands the task from answering a question to maintaining software. Cloud hardware carries the strongest agents today; local agents become the next ownership step as models, context handling, tools, and hardware improve.



The super-intelligence era: models compete on DeepSWE

113 repository-scale tasks · model-vs-model benchmark · beyond practical human completion

1M in + 1M out · CAD/workload · API list price · expensive → cheap



filled circles: published DeepSWE results · hollow circles: inferred from SWE-bench Verified · CAD

Flash-Next Coder derivative: 75.6% SWE-bench Verified · 86.28 LiveCodeBench v6 · no published DeepSWE result

A coding agent inside a real repository faces the kind of work DeepSWE measures: inspecting code, using tools, editing files, and leaving a patch that passes a clean verifier across repository-scale tasks. The points show a steep capability ladder from frontier systems to cheaper open models. A municipality can turn that benchmark into a local test: choose one bounded repair, run the project tests, inspect the patch, and give an accountable person the approval.

More software is made; much stays private

Public code and releases show only part of the output.

RANDOMIZED FIELD TEST
4,867 developers
· 3 companies
+26%
completed tasks
Microsoft · Accenture · Fortune 100

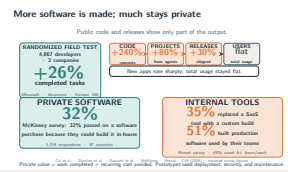
CODE	PROJECTS	RELEASES	USERS
+240%	+80%	+30%	flat
commits	from agents	shipped	total usage

New apps rose sharply; total usage stayed flat.

PRIVATE SOFTWARE
32%
McKinsey survey: 32% passed on a software purchase because they could build it in-house
1,719 respondents · 97 countries

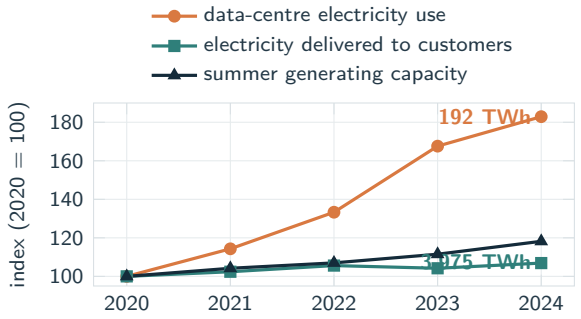
INTERNAL TOOLS
35% replaced a SaaS tool with a custom build
51% built production software used by their teams
Retool survey · 49% saved 6+ hours/week

Cui et al. · Demirer et al. · Daniotti et al. · McKinsey · Retool · CHI (2026) · reported survey figures



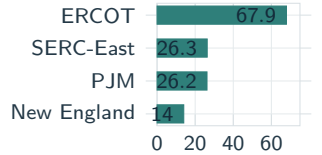
Two measures belong together. Public studies show more code, completed tasks, projects, and releases, with smaller gains as work reaches integration and adoption. Private tools add value through a saved subscription and a workflow shaped around local records. The same coding capability can create a public product or a quiet tool for one town office. In both cases, people still test, release, maintain, and decide whether the result deserves continued use.

Data centres are growing faster than the U.S. grid



+83% data centres **+7%** customer electricity
+18% summer capacity

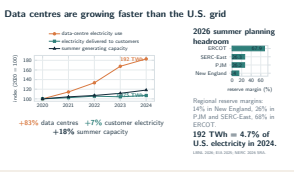
2026 summer planning headroom



Regional reserve margins: 14% in New England, 26% in PJM and SERC-East, 68% in ERCOT.

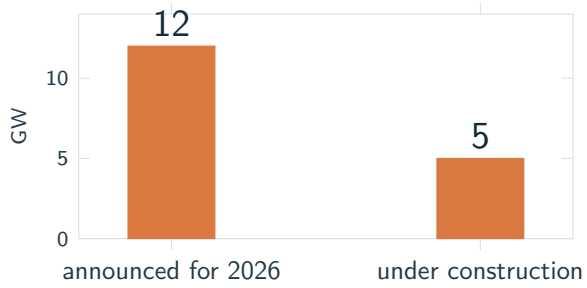
192 TWh = 4.7% of U.S. electricity in 2024.

LBNL 2026; EIA 2025; NERC 2026 SRA.



A new model-serving building waits for its grid connection. U.S. data-centre electricity use has risen faster than customer electricity and summer generating capacity. Each facility needs transmission, a substation, transformers, and a firm connection, while regional reserve margins vary widely. The physical grid sets the pace for new model capacity. That makes the announced projects worth examining as plans for a system with hard construction limits.

The announced data-centre build-out is outrunning construction



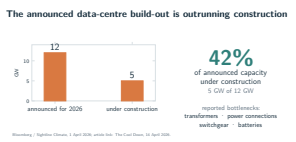
42%

of announced capacity under construction

5 GW of 12 GW

reported bottlenecks:
transformers · power connections
switchgear · batteries

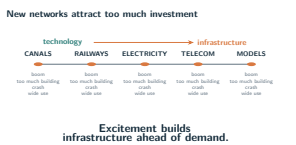
Bloomberg / Sightline Climate, 1 April 2026; article link: The Cool Down, 14 April 2026.



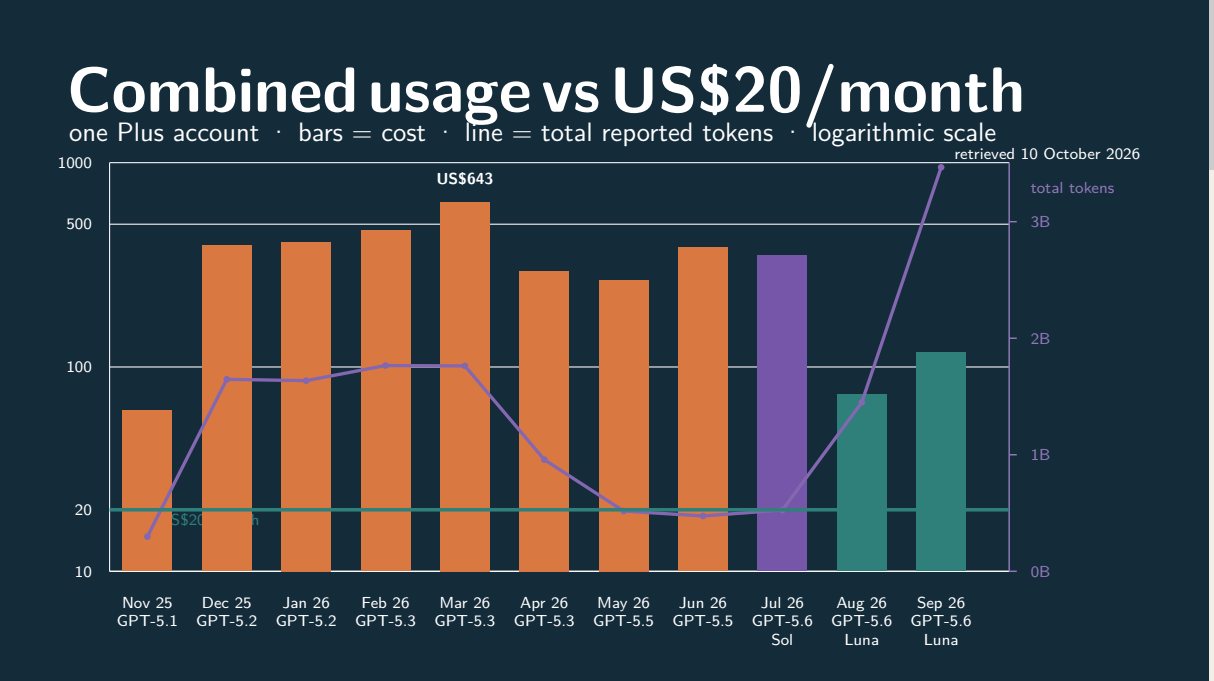
The chart compares announced capacity with construction. Less than half of the announced capacity had reached construction in the survey. Delivery depends on transformers, grid connections, switchgear, batteries, financing, equipment, and committed customers. The gap shows a physical bottleneck: financial announcements can multiply faster than power, machinery, and buildings can be delivered. That sets up the correction that follows when capital outruns demand.

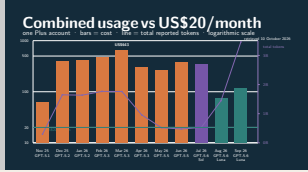
New networks attract too much investment

Excitement builds infrastructure ahead of demand.

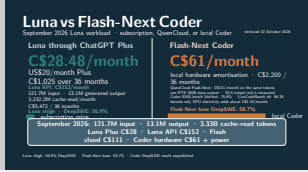


A railway or fibre network can remain after its investors lose money. The tracks and cables remain, and new owners can put them to work at a lower price. Canals, electricity, and communications networks followed this pattern during earlier investment cycles. A model-serving correction could leave racks, fibres, substations, and mature models available to new owners. That is how a national overbuild can become useful computing for a town, farm, school, or workshop.





The industrial story comes down to one account. Each bar shows the hosted-equivalent cost for a full month, and the line shows total input and output tokens, including the cache work recorded by the tools. Heavy months carry much more metered work than the flat subscription suggests; lighter months carry less. October is excluded because it is partial. The chart therefore compares one combined account, one complete month at a time, with the workload visible beside the price.



Luna vs Flash-Next Coder

September 2026 Luna workload · subscription, QwenCloud, or local Coder

retrieved 10 October 2026

Luna through ChatGPT Plus

C$28.48/month

US$20/month Plus

C$1,025 over 36 months

Luna API: C$152/month

121.7M input · 13.1M generated output

3,332.2M cache-read/month

C$5,472 / 36 months

Luna xhigh · DeepSWE: 56.9%

subscription price

Flash-Next Coder

C$61/month

local hardware amortisation · C$2,200 / 36 months

QwenCloud Flash-Next: C$111/month on the same tokens

one RTX 3090-class system · 93.5 output tok/s measured

Coder SWE-bench Verified: 75.6% · LiveCodeBench v6: 86.28

decode-only GPU electricity adds about C$1.35/month

Flash-Next base DeepSWE: 58.7%

local Coder

September 2026: 121.7M input · 13.1M output · 3.33B cache-read tokens

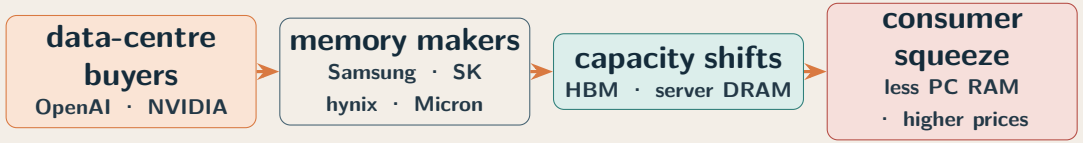
Luna Plus C$28 · Luna API C$152 · Flash

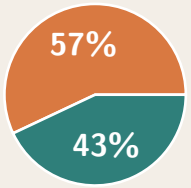
cloud C$111 · Coder hardware C$61 + power

Luna xhigh: 56.9% DeepSWE · Flash-Next base: 58.7% · Coder DeepSWE result unpublished

The same September workload sits beside three paths. The Plus subscription is C$28.48, the Luna API is C$152, QwenCloud Flash is C$111, and a one-card Coder system amortises to about C$61 before full system power. Flash-Next's base model scores 58.7% on DeepSWE, close to Luna xhigh at 56.9%. The compressed Coder has published SWE-bench and LiveCodeBench results; its DeepSWE result remains unpublished, so the cost comparison keeps that boundary visible.

Data-centre demand raises local hardware costs





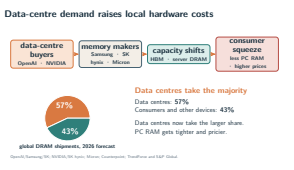
global DRAM shipments, 2026 forecast

Data centres take the majority

Data centres: **57%**
Consumers and other devices: **43%**

Data centres now take the larger share.
PC RAM gets tighter and pricier.

OpenAI/Samsung/SK; NVIDIA/SK hynix; Micron; Counterpoint; TrendForce and S&P Global.



Memory production is moving toward data-centre systems. The forecast assigns a majority of shipped memory to those systems, leaving consumers and local-compute buyers with the smaller share. Large buyers' procurement targets reinforce that demand. Local computers therefore begin with a higher hardware bill. A later correction in data-centre demand could release used memory and accelerators into the local ownership path.

Cloud model services are cheap today for a reason

strategic price
flat subscriptions
pooled demand
venture and infras-
tructure capital

heavy users get leverage
a $20 plan can carry
much more metered value
than its sticker price

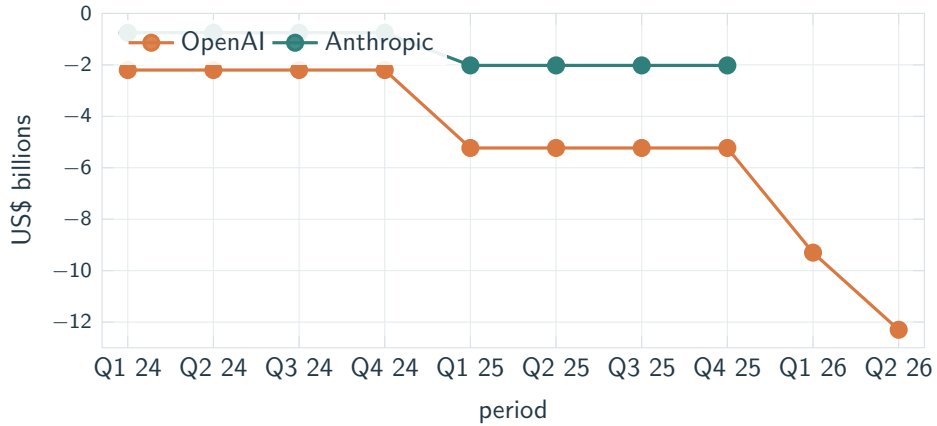
when the bill comes due
higher rates
tighter limits
local looks
better

Subsidy is a market condition that can change.

The slide describes pooled demand and strategic pricing; subscriber-level economics require separate audited data.

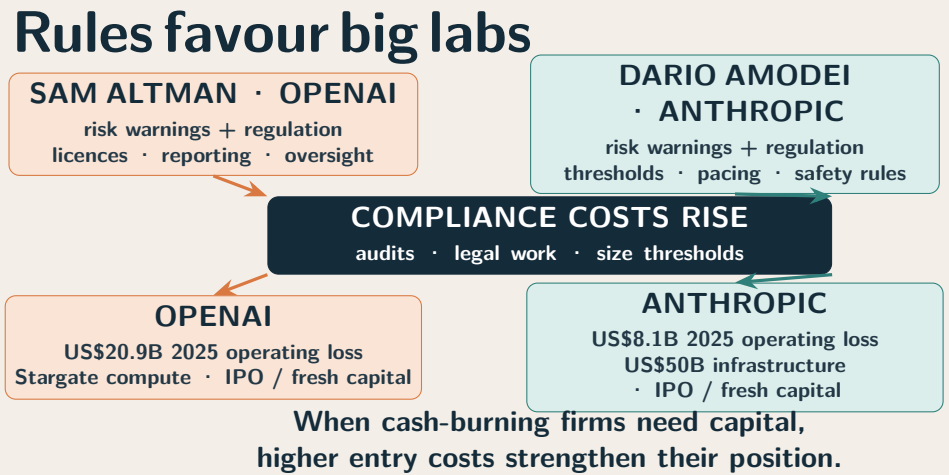
A heavy user can receive so much work for a flat monthly fee because providers pool light and heavy users, raise capital, and build capacity ahead of demand. Each customer's payment covers only part of the system. That arrangement keeps today's price low while the provider carries the cost of chips, power, staff, and unused capacity. A price increase or tighter limit would move the decision toward local ownership.

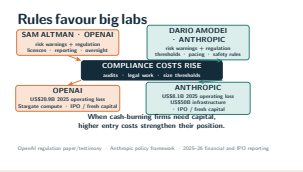
Frontier lab operating losses, quarter by quarter



2024–25 annual losses divided evenly across quarters; 2026 points are reported quarters. Anthropic's adjusted-profit claim is excluded.

The quarterly lines sit below zero. The frontier labs spend heavily on model training, data-centre capacity, staff, and customer service while their prices remain low. Earlier points use a consistent comparison where quarterly figures were unavailable; later points use reported figures. The losses explain the search for outside capital, large infrastructure partners, and rules that make the market easier for firms already operating at this scale.





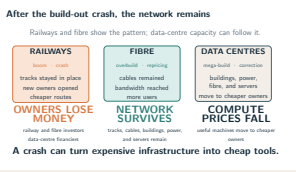
The policy statements sit beside the financing need. OpenAI and Anthropic call for licensing, reporting, thresholds, oversight, and limits around frontier systems. Each requirement can add audits, legal work, reporting, and minimum scale to the cost of entry. When large firms need capital, rules with high fixed costs strengthen firms already able to finance frontier systems. The chart makes that economic effect part of the policy discussion.

After the build-out crash, the network remains

Railways and fibre show the pattern; data-centre capacity can follow it.

RAILWAYS	FIBRE	DATA CENTRES
boom · crash	overbuild · repricing	mega-build · correction
tracks stayed in place new owners opened cheaper routes	cables remained bandwidth reached more users	buildings, power, fibre, and servers move to cheaper owners
OWNERS LOSE MONEY	**NETWORK SURVIVES**	**COMPUTE PRICES FALL**
railway and fibre investors data-centre financiers	tracks, cables, buildings, power, and servers remain	useful machines move to cheaper owners

A crash can turn expensive infrastructure into cheap tools.



The hopeful part of a build-out correction is physical. Railway booms left tracks that later carried cheaper traffic. Fibre overbuilding left cables and bandwidth that reached more users after prices fell. A data-centre correction can leave buildings, power connections, fibre, racks, and mature models in new hands. Investors can lose money while the network survives and useful machines move to cheaper owners. Lower asset prices can widen access to local and regional computing. A crash can turn expensive infrastructure into cheap tools.

A hardware correction can make local coding affordable

Model-system planning midpoints
Qwen3.5:9B C$950 Qwen3.8:27B C$2,200 Flash-Next Coder C$2,200

50% liquidation case capital cost per token roughly halves; throughput remains; power, cooling, maintenance, and staff remain

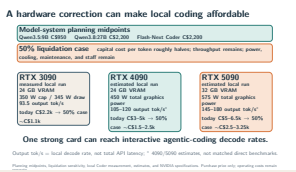
RTX 3090
measured local run
24 GB VRAM
350 W cap / 345 W draw
93.5 output tok/s
today C$2.2k → 50% case ∼C$1.1k

RTX 4090
estimated local run
24 GB VRAM
450 W total graphics power
105–120 output tok/s*
today C$3–5k → 50% case ∼C$1.5–2.5k

RTX 5090
estimated local run
32 GB VRAM
575 W total graphics power
145–180 output tok/s*
today C$5–6.5k → 50% case ∼C$2.5–3.25k

One strong card can reach interactive agentic-coding decode rates.

Output tok/s = local decode rate, not total API latency; * 4090/5090 estimates, not matched direct benchmarks.



This combines the correction economics with the hardware comparison. The cards show today's planning prices beside a simple 50 percent liquidation case: about C$1.1k for the used 3090-class system, C$1.5k to C$2.5k for a 4090, and C$2.5k to C$3.25k for a 5090. These are purchase-price planning sensitivities; market quotations can differ. A liquidation case roughly halves capital cost per token while throughput remains; power, cooling, networking, maintenance, and staff remain operating costs. The RTX 3090 result is measured at 93.5 output tokens per second with the Coder loaded. The 4090 and 5090 figures are estimates, so the slide separates the measured result from the planning cases.

The opportunity: resident-owned local software

WHAT RESIDENTS GAIN

LOWER FEES
keep subscription money in the community

OWNERSHIP
municipality owns the source; residents can inspect it

REPAIR
fix bugs and add features locally

SHARE
publish useful tools for the county and province

MUNICIPAL STARTING POINTS

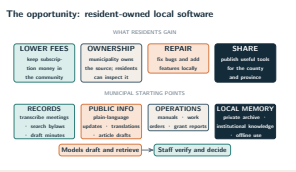
RECORDS
transcribe meetings · search bylaws · draft minutes

PUBLIC INFO
plain-language updates · translations · article drafts

OPERATIONS
manuals · work orders · grant reports

LOCAL MEMORY
private archive · institutional knowledge · offline use

Models draft and retrieve → **Staff verify and decide**



This is the resident payoff and the municipal starting point in one picture. Local ownership keeps subscription money in the community, gives the municipality control of the source and records, and lets a local coding agent repair features or add improvements. The first uses are records, public information, operations, and institutional memory. Models draft and retrieve; staff verify and decide. A town can begin with time-saving work, keep public control, and share a useful tool with the county and province.

Model provenance shapes behaviour

Imported weights carry training choices, policy choices, and context sensitivity.

CONTEXT

DeepSeek-R1

vulnerable output
19% baseline
27.2% in one political / industrial-control test

PERSONA

Qwen family

more code flaws
under a U.S.-government persona in one controlled test

LOCAL FIT

Canadian evaluation

law · history · records
security requirements
public accountability

We evaluate the weights, the prompts, and the complete working system.

CrowdStrike (2025); Booz Allen (2026); NIST/CAISI (2025). Controlled tests measure behaviour under specified conditions.

Model provenance shapes behaviour
Imported weights carry training choices, policy choices, and context sensitivity.
CONTEXT DeepSeek-R1 vulnerable output 19% baseline 27.2% in one political / industrial-control test
PERSONA Qwen family more code flaws under a U.S.-government persona in one controlled test
LOCAL FIT Canadian evaluation law · history · records security requirements public accountability
We evaluate the weights, the prompts, and the complete working system.
CrowdStrike (2025); Booz Allen (2026); NIST/CAISI (2025). Controlled tests measure behaviour under specified conditions.

A small model brings more than general language ability into a public system. Training records, policy rules, safety tuning, and context shape its behaviour. CrowdStrike found vulnerable code in 19 percent of DeepSeek-R1 baseline outputs and 27.2 percent in a political and industrial-control setting. Booz Allen found more Qwen code flaws under a U.S.-government persona. NIST and CAISI found strong differences among models on credential theft, phishing, malware, censorship, and refusal tests. These results show context-sensitive behaviour under test conditions. Canadian systems therefore need Canadian law, records, security requirements, and public accountability in their evaluation.

Independent models can check one another

Different training contexts can expose different blind spots.

DRAFT
one model writes
explicit requirements
security brief
tests to satisfy

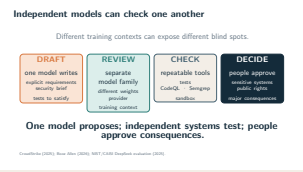
REVIEW
separate
model family
different weights
provider
training context

CHECK
repeatable tools
tests
CodeQL · Semgrep
sandbox

DECIDE
people approve
sensitive systems
public rights
major consequences

One model proposes; independent systems test; people approve consequences.

CrowdStrike (2025); Booz Allen (2026); NIST/CAISI DeepSeek evaluation (2025).



While Canadian models are being trained and evaluated, independent checks can work together. One model writes from a security brief covering authentication, input validation, secrets, dependencies, and tests. A separate model family reviews the result. CodeQL, Semgrep, sandbox tests, and integration tests add repeatable checks. A person approves changes affecting sensitive systems, public rights, or major consequences. Different weights, providers, prompts, and training contexts expose different blind spots. This gives municipalities a practical bridge toward Canadian refinement: diverse checks now, Canadian law and records in the training and evaluation later.

The general-purpose 9B frontier is near today's ceiling

Distillation can move a skill; capacity still sets the range of general work.

Qwen3.5-9B
81.7% GPQA Diamond

Qwen3.8-9B distill
same 9B base; teacher traces from Qwen3.8 2.4T
MMLU +20.5 points GSM8K −1.5 points

FACTUALITY CHECK (5 samples)
matched Qwen3.5-9B on truthfulness

A focused 9B can still climb

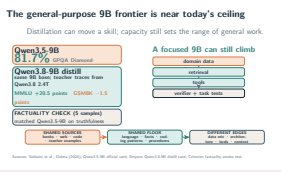
domain data
retrieval
tools
verifier + task tests

SHARED SOURCES
books · web · code · teacher examples

SHARED FLOOR
language · facts · coding patterns · procedures

DIFFERENT EDGES
data mix · architecture · tools · context

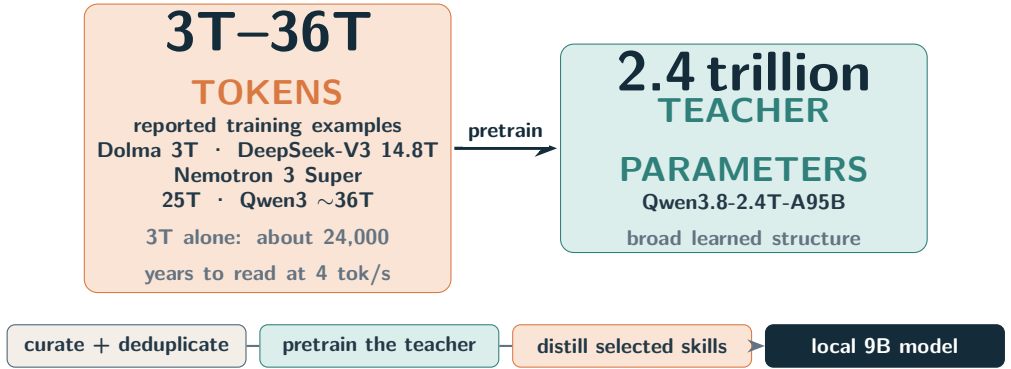
Sources: Soldaini et al.; Dolma (2024); Qwen3.5-9B official card; Empero Qwen3.8-9B distill card; Criterion factuality smoke test.



Models often agree on familiar work because their training records overlap. Books, web pages, code, reference works, and teacher examples create a shared floor of language, common facts, coding patterns, and ordinary procedures. Their differences appear at the edges: data mixture, architecture, post-training, context handling, and tools. The 9B results show the same boundary. The Qwen3.8-9B distill gains 20.5 points on MMLU and gives up 1.5 points on GSM8K under one shared harness; its five-sample factuality check matched Qwen3.5:9B on truthfulness. A focused local system can use the shared floor, then add local records, retrieval, tools, verifiers, and task tests for the work residents actually need.

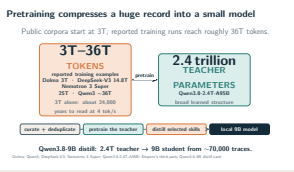
Pretraining compresses a huge record into a small model

Public corpora start at 3T; reported training runs reach roughly 36T tokens.



Qwen3.8-9B distill: 2.4T teacher → 9B student from ∼70,000 traces.

Dolma; Qwen3; DeepSeek-V3; Nemotron 3 Super; Qwen3.8-2.4T-A95B; Empero's third-party Qwen3.8-9B distill card.



The local model is the last step in a much larger process. Dolma's three trillion tokens are one public corpus example. Reported training volumes include 14.8 trillion tokens for DeepSeek-V3, 25 trillion for the Nemotron 3 Super curriculum, and roughly 36 trillion for Qwen3. Three trillion tokens alone would take about 24,000 years to read at four tokens per second. The 2.4 trillion figure on the right is a teacher's parameter count, a different quantity from training tokens. Distillation selects useful behaviour and places it in a smaller model that can run locally.

Large open teachers live in the 400B–3T range

Total parameters include every expert; active parameters are used for each token.

model	total	active/token	BF16 weights	raw 4-bit weights
Qwen3.5-397B-A17B	397B	17B	0.79 TB	0.20 TB
DeepSeek-V3	671B	37B	1.34 TB	0.34 TB
DeepSeek-V4 Pro	1.6T	49B	3.2 TB	0.8 TB
Qwen3.8-2.4T-A95B	2.4T	95B	4.8 TB	1.2 TB
Kimi K3	2.8T	104B	5.6 TB	1.4 TB

A 2.4T teacher needs roughly 38 × 32GB cards for raw 4-bit weights alone.
Runtime memory, context, and communication require additional capacity.

Large open teachers live in the 400B–3T range
Total parameters include every expert; active parameters are used for each token.
model | total | active/token | BF16 weights | raw 4-bit weights
Qwen3.5-397B-A17B | 397B | 17B | 0.79 TB | 0.20 TB
DeepSeek-V3 | 671B | 37B | 1.34 TB | 0.34 TB
DeepSeek-V4 Pro | 1.6T | 49B | 3.2 TB | 0.8 TB
Qwen3.8-2.4T-A95B | 2.4T | 95B | 4.8 TB | 1.2 TB
Kimi K3 | 2.8T | 104B | 5.6 TB | 1.4 TB
A 2.4T teacher needs roughly 38 × 32GB cards for raw 4-bit weights alone.
Runtime memory, context, and communication require additional capacity.

The large open teachers behind many smaller releases sit in a different class from a local 9B or 27B model. Current examples span about 400 billion to 2.8 trillion total parameters. Mixture-of-experts models activate only part of that set for each token, while the full set still has to be stored. Four-bit compression reduces storage, but a 2.4-trillion model still needs about 1.2 terabytes for raw weights alone.

Making a large teacher takes a cluster

DeepSeek-V3 published run

671B total parameters
14.8T pretraining tokens

2.788M H800 GPU-hours

full training run

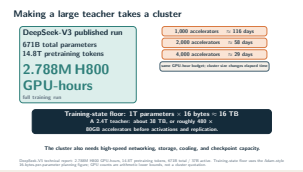
1,000 accelerators	≈ 116 days
2,000 accelerators	≈ 58 days
4,000 accelerators	≈ 29 days

same GPU-hour budget; cluster size changes elapsed time

Training-state floor: 1T parameters × 16 bytes ≈ 16 TB
A 2.4T teacher: about 38 TB, or roughly 480 ×
80GB accelerators before activations and replication.

The cluster also needs high-speed networking, storage, cooling, and checkpoint capacity.

DeepSeek-V3 technical report: 2.788M H800 GPU-hours, 14.8T pretraining tokens, 671B total / 37B active. Training-state floor uses the Adam-style



This is the physical divide in the story. DeepSeek published a full training budget of 2.788 million H800 GPU-hours. Spread over one thousand accelerators, that is about four months; spread over four thousand, about one month. The memory floor grows with every parameter because training keeps weights, gradients, and optimizer state. A 2.4-trillion teacher therefore needs hundreds of 80GB accelerators just for those states, with networking, storage, cooling, and checkpoints around them. The small local model is the product of this concentrated stage.

Quarterly refinements fit a regional model lab

The release cadence changes the building size.

Storage: 3T = 4.5 TB; 14.8T = 22 TB; 32T+ = 48 TB.

LOCAL RE-FINEMENT

8 × 80 GB accelerators

10–15 kW facility

capital ~C$0.5–0.8M

100 TB storage

9B / 27B students

adapters, tests, and domain updates

REGIONAL DISTILLATION

32 × 80 GB accelerators

40–60 kW facility

capital ~C$2–3M

250 TB storage

teacher traces + students

several domain releases / year

FRONTIER TEACHER

1,024–2,048 GPUs

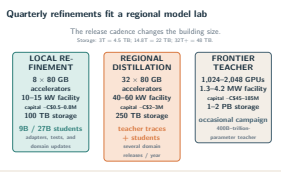
1.3–4.2 MW facility

capital ~C$45–185M

1–2 PB storage

occasional campaign

400B–trillion-parameter teacher



Quarterly is a useful planning cadence for student releases, adapters, domain data, and evaluation. A local refinement room uses 8 accelerators, about 100 terabytes of storage, and C$0.5–0.8 million. A regional lab uses 32 accelerators and C$2–3 million to keep a teacher available and make several releases each year. A frontier campaign uses 1,000–2,000 accelerators and about C$45–185 million. Dolma's 3-trillion-token corpus occupies about 4.5 terabytes compressed; larger training records reach tens of terabytes before copies, checkpoints, and model weights.

Where each computing scale fits in Canada

The lab size is set first; the site must then supply power, fibre, land, and cooling.

Water uses a Canadian cold-climate planning factor of 0.3 L/kWh and is shown per day at midpoint load.

LOCAL RE-FINEMENT

Grey County / Owen Sound

~90 L/day water

Owen Sound industrial lands

existing building · local staff · fibre

one municipal or county lab

REGIONAL DISTILLATION

Montréal · Toronto · Vancouver

~360 L/day water

fibre

3-phase service · technicians

several domain re-leases each year

FRONTIER CAMPAIGN

Québec City–Lévis

Winnipeg · Prince George

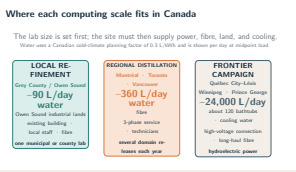
~24,000 L/day

about 120 bathtubs

· cooling water

high-voltage connection · long-haul fibre

hydroelectric power



Site choice follows scale. Grey County and Owen Sound fit a local refinement room using an existing building, local staff, and available fibre; the planning water figure is about 90 litres per day, roughly half a bathtub. Montréal, Toronto, and Vancouver fit regional distillation with fibre, technical staff, customers, and three-phase service; the figure is about 360 litres per day, roughly two bathtubs. Frontier campaigns fit power-rich corridors such as Québec City–Lévis, Winnipeg, and Prince George, with about 24,000 litres per day, roughly 120 bathtubs. Grey-Bruce fits local and regional work; hydroelectric corridors fit frontier work.

Three model-building scales

Residents use the finished models through desktop inference.

LOCAL RE-FINEMENT

municipal model

local bylaws · records · procedures

services fitted to one town

REGIONAL DISTILLATION

province-wide model

provincial law · history · practice shared across municipalities

easier provincial services in every town

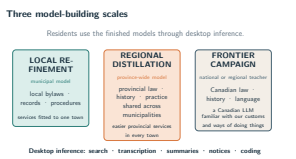
FRONTIER CAMPAIGN

national or regional teacher

Canadian law · history · language

a Canadian LLM familiar with our customs and ways of doing things

Desktop inference: search · transcription · summaries · notices · coding



The layers stay separate. Desktop inference runs the actual services: search, transcription, summaries, notices, and coding. Local refinement produces a municipal model fitted to one town's bylaws, records, and procedures. Regional distillation produces a province-wide model with provincial law, history, and practice; the resident-facing goal is easier navigation of provincial services in every town. A frontier campaign produces a national or regional Canadian teacher familiar with Canadian law, history, language, and customs. The resident-facing goal is a Canadian LLM that understands our customs and ways of doing things.

A larger teacher can produce a fast local student

QWEN3.8 FLASH-NEXT: TEACHER

Conventional GPU server

teacher examples + tests
Canadian records + workflows

refinement run at a regional

lab or frontier data centre

TAALAS HC1: LO-CAL STUDENT

Qwen3.5:9B

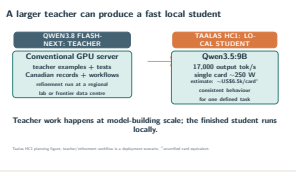
17,000 output tok/s
single card ∼250 W
estimate: ∼US$6.5k/card*

consistent behaviour

for one defined task

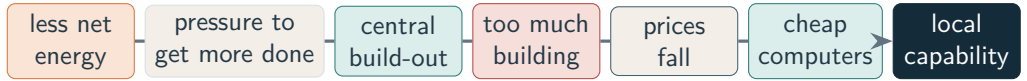
Teacher work happens at model-building scale; the finished student runs locally.

Taalas HC1 planning figure; teacher/refinement workflow is a deployment scenario; *unverified card equivalent.



This slide makes the production chain explicit. Qwen3.8 Flash-Next serves as the teacher on a conventional GPU server. Canadian records, examples, corrections, and tests shape a refinement run at a regional or frontier lab. The result is a Qwen3.5:9B student for one defined task. A Taalas-style 250-watt card can then serve that student at about 17,000 output tokens per second. The teacher stage belongs to model-building infrastructure; the finished student belongs in a municipality, county, or other local setting.

Less surplus energy, more local skill



From less surplus to local skills and production

Global specialisation ⟶ local skills and production

Less surplus energy, more local skill
From less surplus to local skills and production
Global specialisation ⟶ local skills and production

The chain now returns to the local table. Declining surplus raises pressure on labour, money, and coordination; capital builds large model-serving systems; a correction can lower the price of used computers and mature models. A town office, farm, or household can then own a capable machine for repair, crop planning, teaching, or records. The choice is where people place that capability and which useful work they strengthen together.

Community safety: better evidence, better response

A public-safety modern software system can organize lawful evidence while people remain responsible for decisions.

RESPONSIBLE RESPONSE

screen calls by risk
civilian · paramedic · police
police focus serious crime

EVIDENCE WORK

search permitted records
build timelines
prepare disclosure

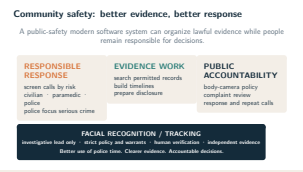
PUBLIC ACCOUNTABILITY

body-camera policy
complaint review
response and repeat calls

FACIAL RECOGNITION / TRACKING

investigative lead only · strict policy and warrants · human verification · independent evidence

Better use of police time. Clearer evidence. Accountable decisions.



Residents want safe public spaces and timely help. A public-safety modern software system can search permitted reports and video records, build a timeline, prepare disclosure, and help match a call to a civilian, paramedic, or police response. Police can keep more time for serious theft, drug trafficking, and emergencies. The same system can publish response times, repeat calls, body-camera policy, and complaint-review results. Facial recognition or continuous tracking carries a stricter boundary: it can serve as an investigative lead under a clear policy or warrant, with human verification and independent evidence before any arrest or conviction. The resident benefit is better use of police time and clearer public accountability.

Healthcare: more help before the clinic

A Canadian public-health modern software system can favour prevention, evidence, and affordable choices.

SELF-CARE	NAVIGATION	PUBLIC CAPACITY
nutrition	provincial services	fewer avoidable visits
prevention	local programmes	less paperwork
evidence in plain language	low-cost options	more time for serious cases

The goal is informed action at low cost, with safety checks for high-risk questions.

Healthcare: more help before the clinic
A Canadian public-health modern software system can favour prevention, evidence, and affordable choices.
SELF-CARE
nutrition
prevention
evidence in plain language
NAVIGATION
provincial services
local programmes
low-cost options
PUBLIC CAPACITY
fewer avoidable visits
less paperwork
more time for serious cases
The goal is informed action at low cost, with safety checks for high-risk questions.

Models inherit the legal, institutional, and commercial setting around their data and deployment. U.S. systems are built inside a for-profit U.S. healthcare market; a Canadian system can set a different public goal. It can help residents understand nutrition and prevention, compare evidence-based low-cost choices, and find provincial or local services before they need a clinical appointment. That can reduce avoidable demand and paperwork while leaving clinicians more time for serious cases. This is a design goal; current models differ by provider, jurisdiction, and deployment.

Education: higher learning close to home

A Canadian teaching modern software system can extend college, graduate, and professional learning.

FOUNDATIONS

teachers build
reading, writing, and numeracy
before tool use

ADVANCED TRAINING

college courses
graduate seminars
professional skills

RESEARCH

literature reviews
methods
original projects

Teachers build the foundation; teaching software can bring graduate-level learning closer to home.

Education: higher learning close to home
A Canadian teaching modern software system can extend college, graduate, and professional learning.
FOUNDATIONS
teachers build
reading, writing, and numeracy
before tool use
ADVANCED TRAINING
college courses
graduate seminars
professional skills
RESEARCH
literature reviews
methods
original projects
Teachers build the foundation; teaching software can bring graduate-level learning closer to home.

Reading, writing, and numeracy are the foundation. Teachers build those abilities and help learners develop judgement before they rely on teaching software. Once that foundation is in place, a Canadian teaching modern software system can provide college courses, graduate-level seminars, professional training, advanced trades instruction, literature reviews, research methods, and help with original projects. Canadian curriculum, history, language, and civic knowledge can shape the examples.

Public services: one clear front door

A civic modern software system can turn scattered rules and forms into a service residents can understand.

FIND	EXPLAIN	ACT
Canadian laws regulations and permits benefits, bylaws, forms	plain language translation accessible answers	navigate applications track requirements keep records

Residents get useful answers with the source document beside them.

Public services: one clear front door
A civic modern software system can turn scattered rules and forms into a service residents can understand.
FIND Canadian laws regulations and permits benefits, bylaws, forms
EXPLAIN plain language translation accessible answers
ACT navigate applications track requirements keep records
Residents get useful answers with the source document beside them.

The same approach applies to municipal, provincial, and federal services. The civic modern software system searches the relevant Canadian laws, regulations, permits, benefits, bylaws, and forms; explains what they require in plain language; and helps residents navigate the application process. Residents keep the source and the record of what they submitted. Municipalities can share open tools with neighbouring towns, the county, and the province, so each office can build on a common service.

ABUNDANT KNOWLEDGE
SCARCE ENERGY
LOCAL CAPABILITY

Pressure creates new levels of co-operation.

Mesopotamia: writing · Sintashta: chariots · Indus: low inequality

A finite world can still be a humane world.
Learn together. Repair together.
Grow together. Care for one another.

Discussion

ABUNDANT KNOWLEDGE
SCARCE ENERGY
LOCAL CAPABILITY
Pressure creates new levels of co-operation.
Mesopotamia: writing · Sintashta: chariots · Indus: low inequality
A finite world can still be a humane world.
Learn together. Repair together.
Grow together. Care for one another.
Discussion

Pressure creates new levels of co-operation. Mesopotamian cities developed writing and records. Sintashta communities developed chariots that made overland travel faster and expanded exchange. Indus cities show large urban communities with little visible wealth concentration. Each example turned shared organization into a lasting human capability. At the kitchen table, a young person repairs a pump, a neighbour brings a meeting recording, and a gardener plans potatoes with local weather and soil records. The computer supplies tutoring, translation, memory, and planning; people supply energy, land, tools, skilled hands, and trust. Shared knowledge keeps useful work close to the people who need it. Learn together, repair together, grow together, and care for one another.

APPENDIX

Evidence, methods, detailed charts, and sources

The main narrative ends here.
Supporting material follows for questions, checking, and discussion.

benchmarks · costs · hardware · energy · sources

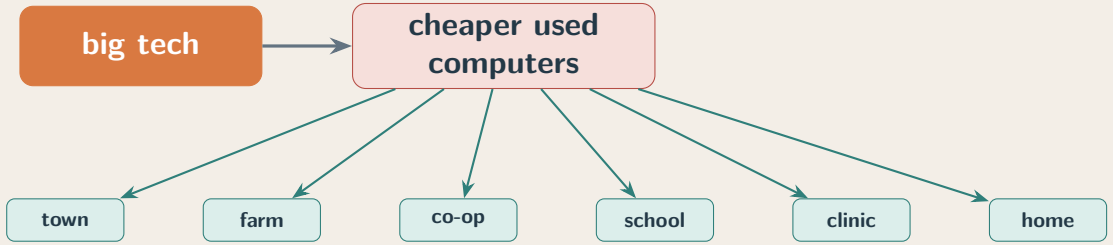
APPENDIX
Evidence, methods, detailed charts, and sources
The main narrative ends here.
Supporting material follows for questions, checking, and discussion.
benchmarks · costs · hardware · energy · sources

The appendix keeps the evidence close at hand. It contains the source figures, calculations, benchmark definitions, and hardware assumptions behind the main story. The main slides use the conclusions; these pages show the measurements behind them. When a number matters, this section gives the scope, the source, and the limit of the comparison. Each detail returns to the local question: which useful task becomes easier to do well?

Who gets the used computers?

used computers become local tools

Who gets the used computers?



A machine built for a large operator can become a local tool after a financing correction. Its value comes from a real workload: records, farming, repair, school, or clinic work. The ownership decision begins with the task, then checks the machine's price, power, throughput, maintenance, and staff time. The question for a town is simple: which used computer can perform one useful job reliably at a full cost the town can carry?

used computers become local tools

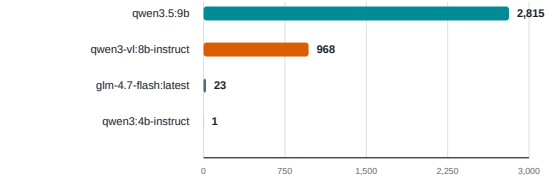
AI After Peak Energy

Ollama did more owned-compute text work than the transcript count suggests

Completed response records in all discovered pyash folders • Qwen3.5:9B is only one part of the owned-compute workload

Completed Ollama responses

Deduplicated by model, created_at, and response counters



OBSERVED OLLAMA VOLUME

3,807	**completed responses**
9.00M	**prompt tokens**
1.32M	**output tokens**
7.46 h	**summed runner time**

1.13 h prompt evaluation • 5.22 h output generation

Qwen3.5:9B alone: 2,815 responses

What the command-bearing files add

Command-bearing files are broader than the deduplicated response total above.

5,723 command-bearing files observed

3,807 distinct responses have counters; 3,154 files have no parsed response metadata

These are Ollama runner records, not ccusage sessions. The summed hours are per-request durations and do not include hardware purchase, electricity, or human review.

Source: all discovered pyash folders, including artifact stores and runtime logs; Ollama response metadata; retrieved 10 October 2026.

The local archive contains more than meeting transcription. Its records include large amounts of text processing, vision work, and smaller amounts of other model use across the pyash trees. Metadata quality varies, so the totals form a documented minimum and a starting point for better recording. The next improvement is simple: record the model, task, token counts, image count, audio length, and energy for every local job.

Paid API equivalent of the owned local workload

Tokens, audio minutes, and image outputs priced in their natural billing units • confirmed records only

Estimated API bill for the same observed work (USD)

orange text • teal transcription • purple image generation

OBSERVED WORKLOAD

6,159 operations

9.00M + 1.32M
input + output tokens

308.84 h
audio represented by transcript timing

2,149 image outputs

Recommended case

$200.66 USD total

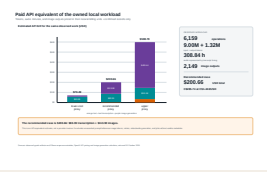
C$285.74 at C$1.424/USD

The recommended case is $200.66: $83.39 transcription + $113.90 images.

This is an API-equivalent estimate, not a provider invoice. It excludes unrecorded prompt/reference-image tokens, retries, video/audio generation, and jobs without usable metadata.

Sources: observed pyash artifacts and Ollama response metadata; OpenAI API pricing and image-generation calculator; retrieved 10 October 2026.

AI After Peak Energy



The local audit can be translated into hosted prices, with separate rates for text, transcription, and image generation. The proxy counts confirmed work and keeps the billing assumptions visible. Local ownership removes API fees from this comparison, while hardware and electricity remain real costs. The result is a fair starting point for asking where local processing saves money and where hosted capacity still makes sense.

Local models move onto owned computers

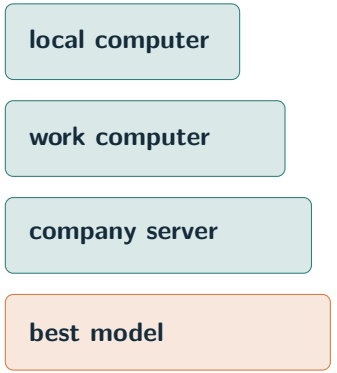
Useful models move from something you rent to something you own.

text · images · audio · video · local files

Local models move onto owned computers
small model
strong local computer
your computer + your data
local workflow
Useful models move from something you rent to something you own.
text · images · audio · video · local files

Local ownership begins with a computer that keeps files nearby. It can transcribe, summarize, search records, and generate media. Serving an existing model requires far less equipment than training a frontier teacher. A workstation can therefore become the first step toward a shared server, with each new task measured by its quality, power use, maintenance, and resident benefit.

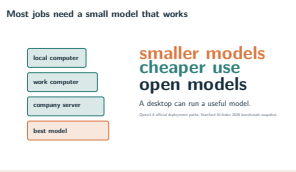
Most jobs need a small model that works



smaller models
cheaper use
open models

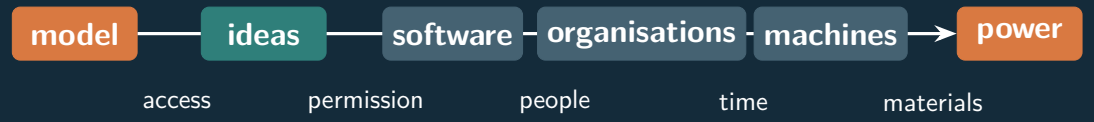
A desktop can run a useful model.

Qwen3.8 official deployment paths; Stanford AI Index 2026 benchmark snapshot.



A small repair shop searches a manual, writes a repair letter, or explains a crop problem. The shop needs dependable answers for those tasks, with the source close to the recommendation. The right measure is good work at a full cost that includes hardware, power, upkeep, and human checking. The same test works for a town office: define the task, keep the evidence beside the answer, and measure whether the result saves time.

Model ideas still need the real world



Models work through the real world.

Model ideas still need the real world

model ideas software organisations machines power
access permission people time materials

Models work through the real world.

A transformer design makes the physical boundary clear. A model can explore options and explain a manual, while mines, factories, crews, materials, power, and time determine what gets built. As the consequence rises, testing, security, and accountable people carry more weight. The model helps with the knowledge work; the community still owns the physical decision and the responsibility for the result.

A model can write. People and tools must check.

Cheap models help people build and check more.

A model can write. People and tools must check.
Cheap models help people build and check more.

A coding agent prepares a change to a public system. The model drafts, tests and scanners find known problems, a separate review looks for missed cases, and a sandbox tries to break the result before approval. The model supplies speed; tools and people supply evidence, judgment, and control. A small municipal team can keep the source, the tests, and the approval record together.

Local production needs knowledge

Local software can bring expert help close to the work

Local production needs knowledge
global system
specialists at the centre
long supply chains
big central institutions
local economy
expertise close to home
repair and reuse
local production
Local software can bring expert help close to the work

A workshop has a pump to repair and a building rule to check. A local software system can search the manual, explain the relevant section, and compare the options with a crop plan or maintenance record. People make the decision and take responsibility for the work. The resident benefit is practical: fewer repeated searches, clearer instructions, and a record that the next worker can use.

A small team can cover more skills

advanced chips
machine tools
parts
repair and software

Local software gives a small operation more skills.

A small team can cover more skills
GLOBAL
CONTINENTAL
NATIONAL
REGIONAL
LOCAL
advanced chips
machine tools
parts
repair and software
Local software gives a small operation more skills.

A replacement part shows the scale difference. Leading-edge chips and large power projects require national or continental systems, while parts, repair, software, food, and many services can move closer to home. Local software lowers the scale at which a small team can afford specialized knowledge. That gives a municipality room to own records, improve its tools, and share the result with neighbouring communities.

Put expert help close to the work

ED

EDUCATION

literacy
numeracy
reasoning
learning on your own

HL

HEALTHCARE

records
monitoring
information
help with decisions

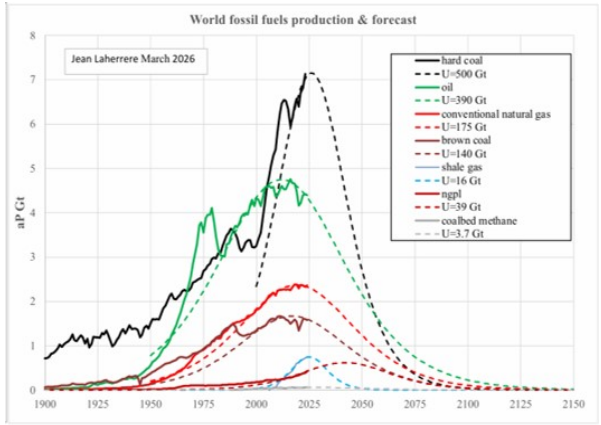
FL

FOOD + LAND

soil
potatoes
food forests
repair
water

Teach people to learn. Give them civilisation's knowledge.

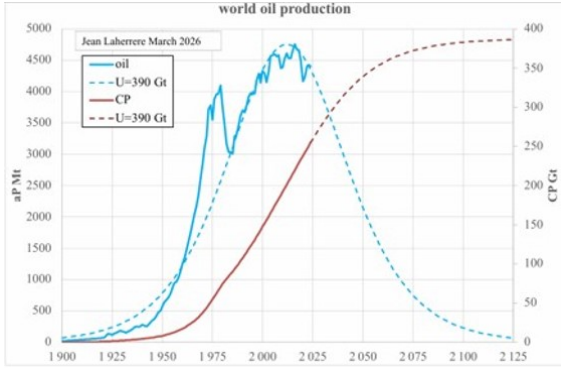
Three mornings show the pattern: a student receives practice, a nurse prepares a clear summary, and a grower compares soil and weather information. Schools, clinicians, and families keep responsibility, while the model supplies explanation, memory, and comparison. The common thread is useful information beside the person doing the job. Local ownership adds continuity: the records, tests, and lessons remain available when staff, students, or service providers change.



Historical Global Production of Fossil Fuels 1900- 2025, and Forecasts to 2100.
U: Estimated ultimately recoverable resource ('URR') in Gt.
Vertical axis. Annual production in Gt.

Source: J. H. Laherrère, *Forecasts of World Fossil Fuel Production to 2150*, ASPO France presentation, 6 August 2026, slide 3. Reproduced from the supplied original.

This source figure supplies the production map behind the opening argument. Its historical curves and forecasts show the sequence of peaks and decline that the main slide summarizes. The detailed assumptions remain here for anyone who wants to examine the forecast. The main point is the direction of the large fuel streams and the pressure that declining useful energy places on transport, food, heat, and public work.

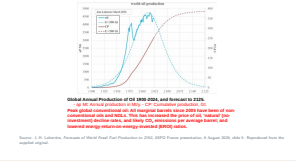


Global Annual Production of Oil 1900-2024, and forecast to 2125.

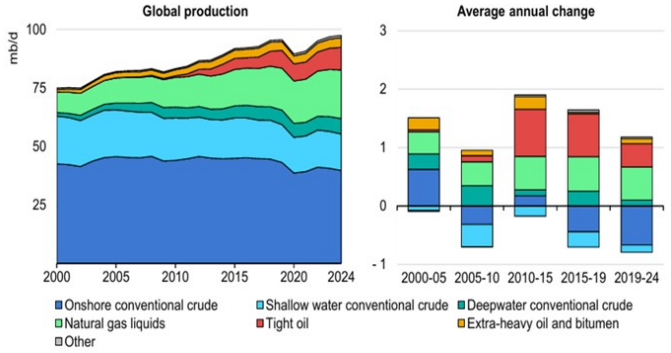
- ap Mt: Annual production in Mt/y. - CP: Cumulative production, Gt.

Peak global conventional oil: All marginal barrels since 2005 have been of non-conventional oils and NGLs. This has increased the price of oil, 'natural' (no-investment) decline rates, and likely CO_2 emissions per average barrel; and lowered energy-return-on-energy-invested (EROI) ratios.

Source: J. H. Laherrère, *Forecasts of World Fossil Fuel Production to 2150*, ASPO France presentation, 6 August 2026, slide 5. Reproduced from the supplied original.



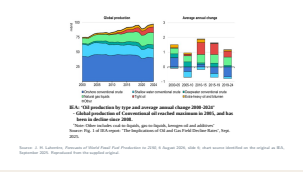
Barrels measure volume, while the work of a town depends on mass, energy, and useful output. Conventional oil has passed its maximum in the figure, and later liquids fill part of the volume with different energy and production burdens. The main argument therefore follows tonnes and useful energy. A full tank can carry a smaller amount of useful work when the fuel stream requires more processing, equipment, and energy to produce.



IEA: 'Oil production by type and average annual change 2000-2024'
- Global production of Conventional oil reached maximum in 2005, and has been in decline since 2008.

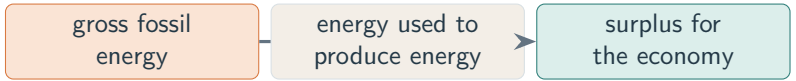
'Note: Other includes coal-to-liquids, gas-to-liquids, kerogen oil and additives'
Source: Fig. 1 of IEA report: 'The Implications of Oil and Gas Field Decline Rates', Sept. 2025.

Source: J. H. Laherrère, *Forecasts of World Fossil Fuel Production to 2150*, 6 August 2026, slide 6; chart source identified on the original as IEA, September 2025. Reproduced from the supplied original.



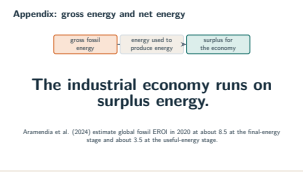
The liquid categories answer different questions. Later additions can keep total volume moving while bringing different masses, energy contents, equipment needs, and processing burdens. Volume and useful energy therefore belong beside separate numbers. For local planning, the useful measure is the energy delivered to transport, buildings, food systems, and public works after the production and conversion costs are paid.

Appendix: gross energy and net energy



The industrial economy runs on surplus energy.

Aramendia et al. (2024) estimate global fossil EROI in 2020 at about 8.5 at the final-energy stage and about 3.5 at the useful-energy stage.



Surplus energy begins with the energy delivered and shrinks through conversion. Final-energy EROI measures energy at delivery; useful-energy EROI follows it into work after conversion losses. For a town, the practical measure is the energy absorbed before transport, food, heating, and public work begin. This is why energy quality and conversion matter alongside the amount of fuel or electricity on the chart.

Appendix: the electricity numbers and their scope

Scope	Source	Now	Forward figure
Global, all data centres	IEA 2026	485 TWh, 2025	945 TWh base case, 2030
United States, all data centres	LBNL 2026	192 TWh, 2024	649 TWh reference case, 2030
Model-serving demand	IEA 2026	grew 50% in 2025	triples, 2025–2030

The global headlines combine multiple loads. Use the denominator for each series.

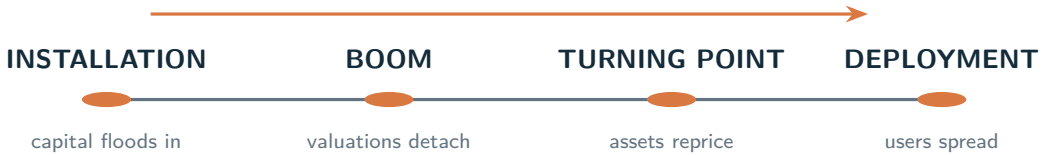
Appendix: the electricity numbers and their scope

Scope | Source | Now | Forward figure
Global, all data centres | IEA 2026 | 485 TWh, 2025 | 945 TWh base case, 2030
United States, all data centres | LBNL 2026 | 192 TWh, 2024 | 649 TWh reference case, 2030
Model-serving demand | IEA 2026 | grew 50% in 2025 | triples, 2025–2030

The global headlines combine multiple loads. Use the denominator for each series.

Every electricity figure needs a boundary. Some sources cover all data centres, and others isolate model-serving equipment. The scope belongs beside the number before demand is compared with local supply. A global all-data-centre forecast answers a different question from the electricity used by model-serving hardware in one country, so the chart keeps those rows separate.

Appendix: installation, crash, deployment

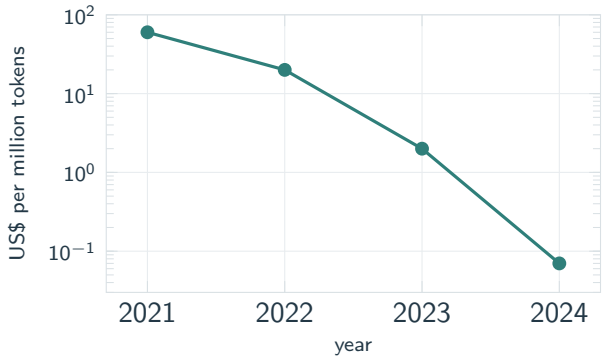


Railways, fibre, and electricity retained value after their financing manias ended.

Appendix: installation, crash, deployment
INSTALLATION BOOM TURNING POINT DEPLOYMENT
capital floods in valuations detach assets reprice users spread
Railways, fibre, and electricity retained value after their financing manias ended.

A financial crash can leave a physical network standing. Canals, railways, electrification, and fibre attracted capital before their full uses appeared; after the break, the network survived and found new owners. Model-serving hardware could follow the same path when mature equipment meets local work. The useful question is who can buy, maintain, and connect those assets after their first owners lose interest.

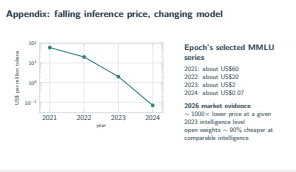
Appendix: falling inference price, changing model



Epoch's selected MMLU series

2021: about US$60
2022: about US$20
2023: about US$2
2024: about US$0.07

2026 market evidence
$\sim 1000\times$ lower price at a given 2023 intelligence level
open weights $\sim$ 90% cheaper at comparable intelligence



The falling price becomes meaningful when the task stays fixed. A given level of text performance costs less over time, while complex tasks consume more energy and compute. The sound comparison holds the task and quality level constant. A town should therefore compare the cost of one defined job, such as searching bylaws or preparing a transcript, at a defined quality level.

Appendix: safety policy and market structure

Document	What it puts on the record
OpenAI, 2023	frontier standards, registration, evaluation, and oversight proposal
OpenAI, 2026	frontier governance framework and risk-management process
Anthropic, 2026	compute, revenue, R&D, and capability thresholds in responsible scaling
FTC, 2025	competition risks in model/cloud partnerships: compute, switching costs, talent, data, control

Public safety rationale and incumbent advantage can coexist in the same policy design.

Appendix: safety policy and market structure

Document | What it puts on the record
OpenAI, 2023 | frontier standards, registration, evaluation, and oversight proposal
OpenAI, 2026 | frontier governance framework and risk-management process
Anthropic, 2026 | compute, revenue, R&D, and capability thresholds in responsible scaling
FTC, 2025 | competition risks in model/cloud partnerships: compute, switching costs, talent, data, control

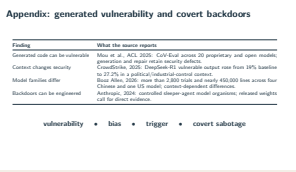
Public safety rationale and incumbent advantage can coexist in the same policy design.

The policy documents list evaluations, thresholds, reporting, oversight, partnerships, and investment. Each item also carries legal work, audits, reporting, capital, and computer time. Those costs matter most to smaller builders. The stated public rationale and the market effect belong in the same discussion when a rule raises the fixed cost of entering the field.

Appendix: generated vulnerability and covert backdoors

Finding	What the source reports
Generated code can be vulnerable	Mou et al., ACL 2025: CoV-Eval across 20 proprietary and open models; generation and repair retain security defects.
Context changes security	CrowdStrike, 2025: DeepSeek-R1 vulnerable output rose from 19% baseline to 27.2% in a political/industrial-control context.
Model families differ	Booz Allen, 2026: more than 2,800 trials and nearly 450,000 lines across four Chinese and one US model; context-dependent differences.
Backdoors can be engineered	Anthropic, 2024: controlled sleeper-agent model organisms; released weights call for direct evidence.

vulnerability • bias • trigger • covert sabotage


Appendix: generated vulnerability and covert backdoors

Finding | What the source reports

Generated code can be vulnerable | Mou et al., ACL 2025: CoV-Eval across 20 proprietary and open models; generation and repair retain security defects.

Context changes security | CrowdStrike, 2025: DeepSeek-R1 vulnerable output rose from 19% baseline to 27.2% in a political/industrial-control context.

Model families differ | Booz Allen, 2026: more than 2,800 trials and nearly 450,000 lines across four Chinese and one US model; context-dependent differences.

Backdoors can be engineered | Anthropic, 2024: controlled sleeper-agent model organisms; released weights call for direct evidence.

vulnerability • bias • trigger • covert sabotage

Generated code can create security failures through different prompts and model families, and controlled work shows that conditional behaviour can be engineered. The response is layered checking: test across model families, sandbox the code, scan it, and require human approval for high-consequence systems. A public system should keep the test results and approval record so later staff can understand why a change was accepted.

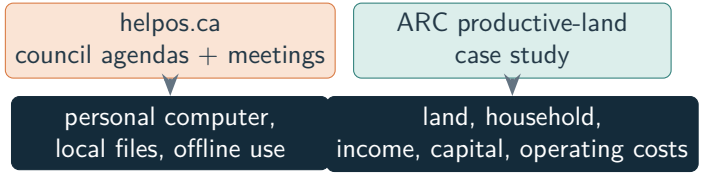
Appendix: local institutions and energy transition

Education	Healthcare	Agriculture
Létourneau et al. (2025) review 28 studies and 4,597 K–12 students. Intelligent tutoring shows generally positive, moderated effects. Physical schools retain social, supervisory, and practical functions.	WHO (2025) identifies retrieval, records summarisation, patient education, monitoring, and decision support as relevant uses, with governance and evidence requirements.	FAO documents energy-intensive fertiliser, machinery, irrigation, processing, and transport. Sánchez et al. (2022) find diversified systems at least as profitable on average, with metric- and context-dependent trade-offs.

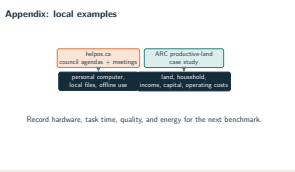
Appendix: local institutions and energy transition

Coordinated software systems meet institutions through ordinary work. Tutoring, health administration, and food planning combine knowledge work with physical systems, care, and responsibility. Each project needs its own safety, labour, and cost measures before people rely on it. The common design is a model connected to records, tools, checks, and people who remain responsible for the outcome.

Appendix: local examples



Record hardware, task time, quality, and energy for the next benchmark.



Helpos turns agendas and meeting recordings into searchable local records. A productive-land project adds land, household, income, capital, and operating-cost assumptions. These examples show two kinds of local work: public records and physical production. A useful benchmark records the task, quality checks, time, and energy so residents can judge the result and improve the system over time.

Appendix: productivity results from current studies

+240%
coding activity
autonomous agents · 500k+ developers
NBER 2026

+26%
tasks completed
4,867 developers · field RCTs
Management Science 2026

>50%
code output
Ant Group CodeFuse
Journal of Financial Stability 2026

55.8%
faster coding task
JavaScript HTTP server
Copilot experiment 2023

40%
less writing time
quality +18% · 453 professionals
Science 2023

+15%
issues per hour
5,172 support agents
QJE 2025

Task outcomes from Demirer et al. (2026), Cui et al. (2026), Gambacorta et al. (2026), Peng et al. (2023), Noy and Zhang (2023), and Brynjolfsson et al. (2025).

More generated work enters a chain that runs from output to integration, release, adoption, and demand. The studies measure different points in that chain, with smaller gains farther downstream. A large increase in code activity can therefore produce a smaller increase in released software. The result people can use belongs beside the amount of model activity, especially for a municipal tool that must keep working over years.

Appendix: learning results from current studies

>2×
learning gains
49 vs 60 minutes · 194 physics students
Scientific Reports 2025

+0.27 SD
knowledge test
gains retained after one week
Contractor and Reyes 2026

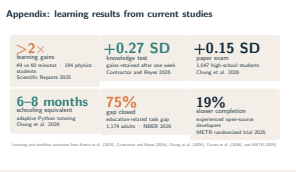
+0.15 SD
paper exam
1,047 high-school students
Chung et al. 2026

6–8 months
schooling equivalent
adaptive Python tutoring
Chung et al. 2026

75%
gap closed
education-related task gap
1,174 adults · NBER 2026

19%
slower completion
experienced open-source developers
METR randomized trial 2025

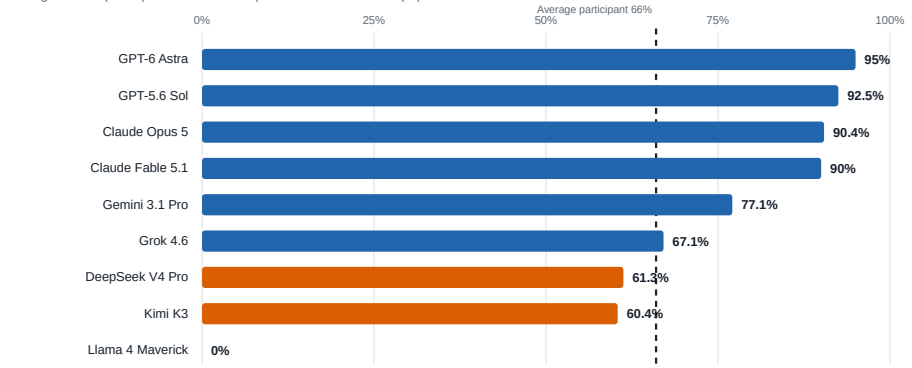
Learning and workflow outcomes from Kestin et al. (2025), Contractor and Reyes (2026), Chung et al. (2026), Cruces et al. (2026), and METR (2025).



Structured tasks and feedback can improve learning, while early coding tools could lengthen experienced developers' work. Task design, user experience, model behaviour, workflow, and human effort shape the result. The useful comparison is a real task with a clear quality check. A municipality should measure time, correctness, review effort, and maintenance after deployment alongside any speed claim.

ARC-AGI-2: model score against human performance

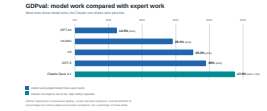
Average human participant: 66% of attempted tasks • ARC-AGI-2 paper



ARC Prize official evaluation data • v2 semi-private set • retrieval 2026-09-19

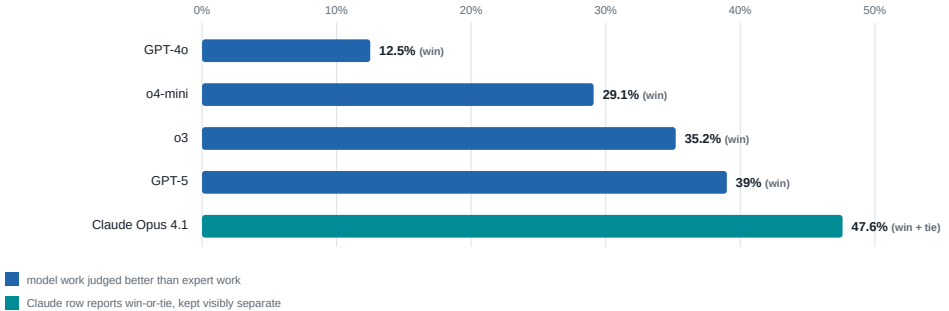
This is a task-specific comparison, not a general intelligence score.

ARC-AGI-2 adds a second capability check. Its bars separate downloadable-weight systems from closed or hosted systems, and its tasks test abstract reasoning. The score gives one comparison point. A town work trial adds the conditions that matter locally: source records, tools, permissions, tests, and an accountable person who decides whether the result is ready.



GDPval: model work compared with expert work

Most rows show model wins; the Claude row shows wins plus ties



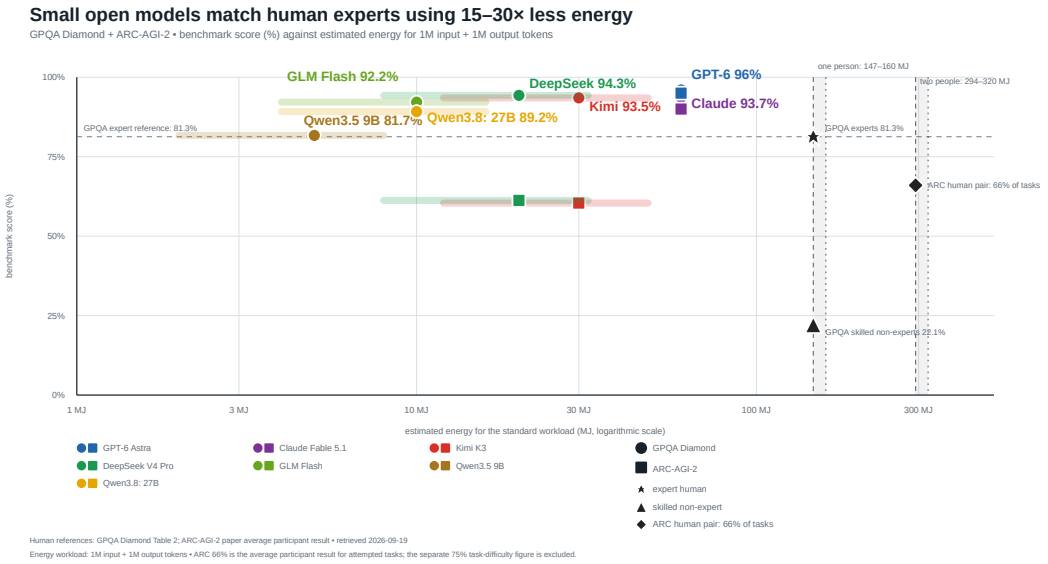
model work judged better than expert work

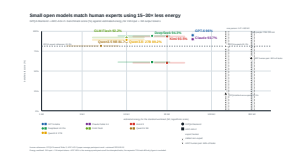
Claude row reports win-or-tie, kept visibly separate

GDPval: OpenAI task set and expert grading • vendor-reported comparison • retrieval 2026-09-19

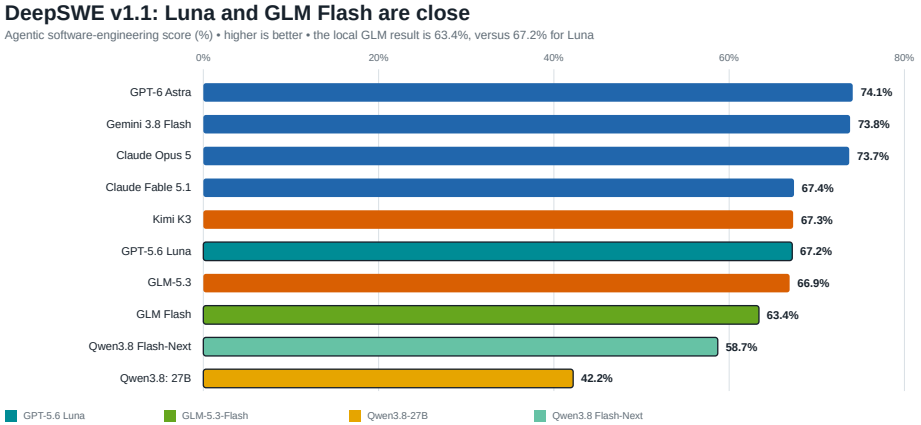
A percentage here means judged work-product comparison, not a percentage of human ability.

GDPval brings the comparison into office work across occupations. It places model deliverables beside expert deliverables. The comparison helps show the range of office work a model can attempt. Residents and staff still judge reliability, review the sources, and decide whether the work remains useful after the first draft.



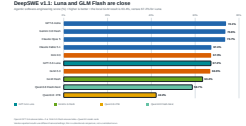


This chart joins capability and energy for a defined workload. GPQA and ARC-AGI-2 measure different abilities, and the horizontal axis estimates energy for common reading and writing. Each benchmark's definition belongs beside its score. For local planning, the useful question is which model clears the required quality level while using a manageable amount of power on the actual job.

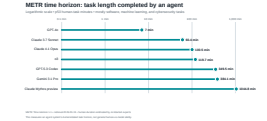


OpenAI GPT-5.6 release table • Z.ai GLM-5.3-Flash release table • Qwen3.8 model cards

Vendor-reported results use different harness/settings; this is a directional comparison, not a controlled rerun.

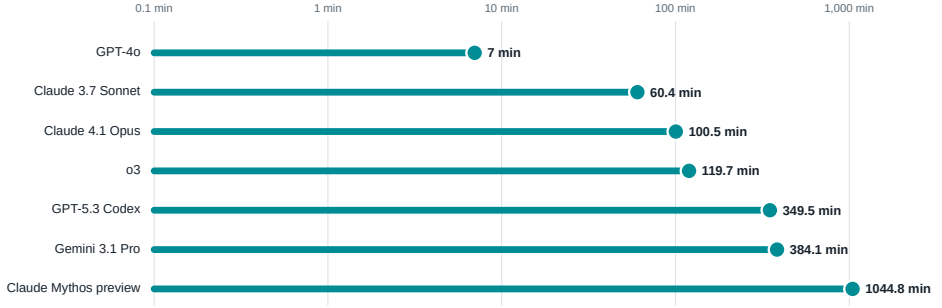


This benchmark gives the Luna and local-hardware discussion a common reference. Frontier systems sit higher, while smaller open models cover bounded work at lower cost. Harnesses and sampling settings differ across rows, so compare capability and cost together. A municipal trial should repeat the comparison on its own repository before a purchase or deployment decision.



METR time horizon: task length completed by an agent

Logarithmic scale • p50 human-task minutes • mostly software, machine learning, and cybersecurity tasks



METR Time Horizon 1.1 • retrieval 2026-09-19 • human duration estimated by contracted experts
This measures an agent system's demonstrated task horizon, not general human-vs-model ability.

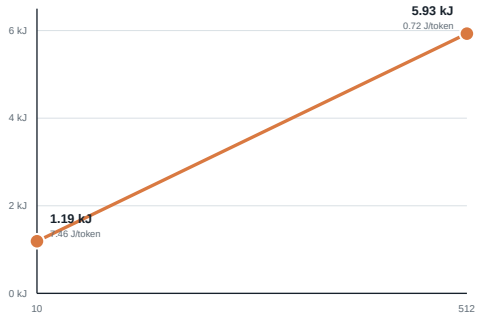
METR measures how long an agent can work independently on technical tasks. It times human experts and asks whether an agent completes tasks of comparable duration. That time horizon complements direct benchmark scores and a local trial. The measure matters for maintenance work because a useful coding agent must carry a task through several steps, from repository inspection to a tested change and review.

Three quantities stay separate: direct energy per generated token, full serving cost, and human labour for the same workload. Each answers a different question. The boundaries stay visible when local machines and hosted services are compared. A fair municipal estimate includes the computer, electricity, maintenance, human checking, and the quality standard required for the job.

Two useful energy comparisons

Provider serving energy and human metabolism answer different questions

Llama 3.2 1B on H200



Total request energy rises

Fixed setup work is spread across more tokens

A provider's cost per token

electricity | hardware | facility | network | operations

internal serving cost

balanced API bill

1M input × US$2 = US$2

1M output × US$10 = US$10

total: US$12

provider serving midpoint: about 1 MJ per US$1

balanced human comparison

read 1M input tokens: 52.5 hours • 57 MJ

generate 1M output tokens: 83.3 hours • 90 MJ

147 MJ ÷ US$12 = 12.25 MJ per US$1

135.8 hours of human work • working comparison

Vellaisamy 2026 • Samsi 2023 • Oviedo 2025 • Liu 2026 • OpenAI and Anthropic pricing

Human time and metabolism use the presentation's balanced comparison; provider serving energy uses measured workload studies.