AI electricity and the work it replaces

Ordinary AI text replies can use little electricity. Large reasoning runs, research agents and generated video can use much more. Whether AI reduces the electricity needed to complete a job depends on the entire workflow, including failed attempts, human review and what the workplace does with the time saved.

Evidence is current to 14 September 2026. The measurement date and system boundary accompany each major figure. The report distinguishes direct measurements, published engineering estimates and editable illustrative scenarios. None of the illustrative scenarios is a measured average for an Ontario household or an AI product.

Main findings

Question	Finding	Evidence or calculation
How much for a text reply?	A useful short-text reference is 0.3 Wh.	Google measured a 0.24 Wh median in May 2025; a June 2026 study models 0.31 Wh for short requests. Neither is a universal current-model average.[2,32]
How much for research or media?	The workload matters more than the word “prompt.”	Measured agent and video workloads span orders of magnitude; hardware-only measurements omit some electricity.[33,36,37]
Does 4 hours becoming 20 minutes save 92% of energy?	Usually not if the building stays on.	Illustrative office: 1.390 to 1.211 kWh over the same 4 hours, with 30 Wh of AI work: 12.9% less.
Can faster AI work use more electricity?	Yes.	With 300 Wh of AI work in that same office example, electricity rises to 1.481 kWh, 6.5% more.
Are data centres environmentally trivial?	No.	IEA estimates 485 TWh in 2025 and projects about 950 TWh in 2030 for all data centres.[1]

The most defensible comparison is electricity per accepted result, assessed over a common period. Time savings are valuable even when they do not switch off a building. Conversely, a low-energy prompt does not erase the impacts of billions of requests, new construction, power generation or local water demand.

How to use the accompanying calculator

The workbook contains editable office, home, project, token-budget and water/carbon calculations, plus source-linked benchmark data. Amber cells are assumptions. Formula cells show the arithmetic. Change the AI energy budget, workstation power, hours, cooling duty cycle and after-work behaviour; the conclusions can change direction.

Source notes

1. IEA. Key Questions on Energy and AI

2. Google researchers. Measuring the environmental impact of delivering AI at Google Scale

32. Felipe Oviedo and colleagues. Energy Use of AI Inference, Efficiency Pathways, and Test-Time Scaling

33. ML.ENERGY research team. Leaderboard task datasets

36. Kim, Shin, Chung and Rhu. The Cost of Dynamic Reasoning: Demystifying the Latency and Energy Overhead of Agentic Workflows

37. Pham and colleagues, Brave Research. AgentStop: Terminating Local AI Agents Early to Save Energy in Consumer Devices

Everyday comparisons with sources

One watt-hour (Wh) is the electricity used by a 1 W device for one hour. One kilowatt-hour (kWh) is 1,000 Wh. A 10 W lamp uses 1 Wh in six minutes. Divide the familiar activity's Wh by the AI workload's Wh to obtain an electricity equivalent.[11]

Activity	Electricity	At 0.3 Wh each	Source and boundary
Full smartphone charge	22.2 Wh	74 short replies	Pixel 9 rated battery 17.77 Wh / assumed 80% charging efficiency. Wall electricity; not a measured charging session.[7]
One hour Netflix	77 Wh	257 short replies	IEA estimate for 2019 streaming, including viewing device and networks; published 2020.[5]
One hour video streaming	188 Wh	627 short replies	Carbon Trust European 2020 benchmark; different device mix and allocation method.[6]
Porch LED overnight	120 Wh	400 short replies	Assumed 10 W bulb x 12 h. Actual wattage and night length determine result.[8,11]
Older porch bulb overnight	720 Wh	2,400 short replies	Assumed 60 W bulb x 12 h.[8,11]
Central AC, modest day	6,000 Wh	20,000 short replies	Assumed 3 kW including blower x 2 equivalent full-load hours; not a Toronto weather prediction.[9,10]

The 0.3 Wh reference is a rounded planning anchor for a short text exchange, informed by production measurement and engineering estimates. It is not an observed average for every present-day model. A 3 Wh request gives one-tenth as many equivalents; a 30 Wh agent job gives one-hundredth. For example, a phone charge equals about 0.74 of a 30 Wh job.[2,3,32]

The two streaming figures are separate historical estimates, not the endpoints of a measured 2026 confidence interval. A television, laptop and phone have different power demands. Do not multiply streamed gigabytes by an old network-intensity number and assume every byte causes that much additional electricity. Much network equipment stays powered regardless of this one stream.[5,6]

For a fully matched activity comparison, add the AI user's device and access network. At an assumed 60 W workstation, two minutes of reading adds 2 Wh: a 0.3 Wh server reply becomes at least 2.3 Wh before network allocation. If the workstation was already in use, its incremental draw may be much smaller. These questions of allocation also apply to streaming.

Source notes

2. Google researchers. Measuring the environmental impact of delivering AI at Google Scale
3. Josh You, Epoch AI. How much energy does ChatGPT use
5. George Kamiya, IEA. The carbon footprint of streaming video fact checking the headlines
6. Carbon Trust. Carbon impact of video streaming
7. Google. Lithium battery test summary
8. US Department of Energy. Purchasing Energy Efficient Light Bulbs
9. Natural Resources Canada. Air Conditioning Your Home
10. Natural Resources Canada. Keeping the Heat In Section 9
11. Virginia Cooperative Extension. Energy Series Estimating Appliance and Home Electronic Energy Use
32. Felipe Oviedo and colleagues. Energy Use of AI Inference, Efficiency Pathways, and Test-Time Scaling

What can be said about current models

There is no sufficiently comparable public electricity dataset to assign a trustworthy average Wh per prompt to every popular current commercial model. Current names and capabilities are public; deployment hardware, batching, utilization, model routing and full request-energy distributions often are not. An energy figure for an older version cannot be transferred to a new release merely because the brand is unchanged.

Product family	Current examples at review date	Defensible energy statement
OpenAI	GPT-6 Astra; GPT-5.6 Sol, Terra and Luna; GPT-Image-2.5 variants	Official catalogue confirms the releases. No directly comparable per-model production Wh averages are established here.[39]
Anthropic	Claude Fable 5.1, Mythos 5.1; Opus 5	Recent release notes establish product versions. Older agent or open-model results do not measure these services.[40]
Google	Gemini 3.8 Flash; Gemini Omni 1.1 Flash video	The 0.24 Wh result is for the median Gemini Apps text request in May 2025, not a measurement of these newer models.[2,41]
Open-weight models	Qwen, DeepSeek, Llama, GPT-OSS; diffusion models	Hardware benchmarks offer repeatable measurements for specified versions and workloads. They do not reveal every cloud provider's energy per request.[33,34]

The strongest short-text anchors

Google's May 2025 production study reports a median of 0.24 Wh for a Gemini Apps text prompt. Its accounting includes accelerators, host systems, idle capacity allocation and facility overhead; active accelerator energy alone was about 0.10 Wh. It excludes training, user devices and external networks. The distinction shows why multiplying GPU energy only by a cooling factor can undercount a service.[2]

Oviedo and colleagues' updated June 2026 model gives a short-request median of 0.31 Wh, with an interquartile range of 0.16-0.60 Wh. Its longer test-time-compute distribution has a median of 3.91 Wh and interquartile range of 2.15-7.05 Wh. These are modeled distributions for throughput-optimized deployments, not meter readings for named current products. Earlier versions of the paper reported different numbers.[32]

A median is the middle request; the arithmetic mean is required to multiply by total request count. Long jobs can pull the mean far above the median. A provider disclosure should identify the mean, median, upper percentiles, request mix, date, included infrastructure and total electricity.

Source notes

2. Google researchers. Measuring the environmental impact of delivering AI at Google Scale
32. Felipe Oviedo and colleagues. Energy Use of AI Inference, Efficiency Pathways, and Test-Time Scaling
33. ML.ENERGY research team. Leaderboard task datasets
34. Jae-Won Chung and colleagues. Where Do the Joules Go? Diagnosing Inference Energy Consumption
39. OpenAI. Models
40. Anthropic. Release notes
41. Google AI for Developers. Gemini API changelog

Measured text and reasoning benchmarks

The following are selected ML.ENERGY measurements, converted from joules to Wh. For each model and task, the lowest-energy B200 configuration in the retrieved dataset is shown. This selection represents efficient tested configurations, not average interactive service operation. GPU energy is amortized across the tested workload; host computers, unassigned idle capacity and facility cooling are excluded.[33,34]

Model / workload	GPU Wh / request	Mean output tokens	GPUs / max sequences
Qwen 3 30B A3B Instruct / chat	0.0145	919	1 / 1,536
Llama 3.3 70B / chat	0.0248	409	2 / 1,024
DeepSeek V3.1 / chat	0.1418	842	8 / 3,072
Qwen 3 30B A3B Thinking / GPQA	0.1982	8,143	1 / 512
Qwen 3 32B / GPQA	0.6090	7,035	1 / 128
DeepSeek V3.1 / GPQA	3.3261	8,969	8 / 256
DeepSeek R1 / GPQA	7.4499	11,287	8 / 128
GPT-OSS 120B / GPQA	0.0198	1,794	1 / 3,072

Source for every row: pinned ML.ENERGY repository snapshot, accessed 14 September 2026. “Max sequences” is the server configuration ceiling, not proof that every batch had that many requests. Chat and GPQA reasoning differ in difficulty, answer length and achieved quality. These rows must not be read as a quality-adjusted product ranking.[33]

Why benchmark numbers can look surprisingly small

Batching shares model-weight movement and infrastructure across many requests. Single-user local inference or a latency-sensitive service may achieve much less sharing. Hardware generation, quantization, cache reuse and speculative decoding also affect energy. An efficient measured configuration may not meet the latency or accuracy requirement of a different job.[34]

The useful result is the range of observed behaviour under disclosed conditions. It is not a claim that one GPT-OSS request is always hundreds of times greener than a DeepSeek request. For deployment decisions, run the same tasks, require the same success criteria and meter the entire serving system.

Illustrative boundary conversion: 1 Wh of GPU energy x 1.25 for host systems x 1.15 for idle-capacity allocation x 1.20 PUE = 1.725 Wh at the facility. All three multipliers are assumptions. Never apply this uplift again to a published number that already includes those components.

Source notes

33. ML.ENERGY research team. Leaderboard task datasets

34. Jae-Won Chung and colleagues. Where Do the Joules Go? Diagnosing Inference Energy Consumption

Research agents and coding agents

An agent is a workflow, not a fixed-sized prompt. It can search, revisit documents, call models repeatedly, launch parallel workers, execute code and retry failures. Final answer length and elapsed minutes do not reveal the amount of computation. Waiting for a web page is different from occupying eight GPUs.

Measured workload	Electricity	What the measurement covers
Llama 3.1 8B, HotpotQA, Reflexion	41.53 Wh / task	One A100; 38% task accuracy; GPU energy only.[36]
Llama 3.1 8B, HotpotQA, LATS	22.76 Wh / task	One A100; 80% accuracy; different search strategy.[36]
Llama 3.1 70B, HotpotQA, Reflexion	348.41 Wh / task	Eight A100s; 67% accuracy; older serving setup.[36]
Llama 3.1 70B, HotpotQA, LATS	158.48 Wh / task	Eight A100s; 82% accuracy. Not a current commercial research-service average.[36]
Local Qwen3-Coder 30B A3B, SWE-bench Verified	2.716 Wh / attempt	Apple M1 Max CPU/GPU telemetry; 500 tasks, 18.8% success; excludes full wall-plug and remote tools.[37]

The A100 experiments isolate agent overhead on specific academic tasks with limited sharing across independent users. Their absolute values should not be projected onto today's highly batched services. They establish that a multi-step agent can cost much more than a single chat turn, and that algorithm choice changes both energy and success.[36]

In AgentStop, the local coding benchmark averaged 1.467 Wh for successful runs and 3.005 Wh for failed runs. Dividing all 500 attempts' approximately 1,358 Wh by 94 successes yields 14.4 Wh per observed success, including failed-attempt electricity. This does not finish the other 406 tasks; human fallback energy remains additional.[37]

A defensible commercial-agent estimate

Obtain a trace of model calls, input and generated tokens, cache hits, retries, tool runtime and media candidates. Apply a workload-specific energy model with explicit uncertainty; add server-side tools and infrastructure only where not already counted. If the trace is unavailable, present a task budget and sensitivity analysis, not an invented average.

Anthropic reported roughly 4 times the tokens for agents and 15 times for multi-agent systems versus chat in its 2025 implementation. Token multipliers are not electricity multipliers: routing, model size, context reuse and serving hardware differ.[38]

Source notes

36. Kim, Shin, Chung and Rhu. The Cost of Dynamic Reasoning: Demystifying the Latency and Energy Overhead of Agentic Workflows
37. Pham and colleagues, Brave Research. AgentStop: Terminating Local AI Agents Early to Save Energy in Consumer Devices
38. Anthropic engineering. How we built our multi-agent research system

Image and video generation

A generated picture is not interchangeable with a text reply. Size, number of denoising steps, architecture, edits, candidate count and upscaling matter. The table shows selected minimum-energy B200 measurements from ML.ENERGY. These are GPU-only figures; different models do not necessarily produce equivalent quality.[33]

Model / output	GPU Wh each	Configuration
SANA 1.5 1.6B / image	0.094	1,024 x 1,024; 20 steps; batch 4
Stable Diffusion 3.5 Medium / image	0.450	1,024 x 1,024; 28 steps; batch 1
Stable Diffusion 3.5 Large / image	0.978	1,024 x 1,024; 28 steps; batch 16
FLUX.1 Dev / image	1.855	1,024 x 1,024; 50 steps; batch 1
CogVideoX 2B / video	7.32	480 x 720; 49 frames at 8 fps; 50 steps
Wan 2.1 1.3B / video	16.91	480 x 832; 81 frames at 15 fps; 50 steps
Wan 2.1 14B / video	79.00	480 x 832; 81 frames at 15 fps; 50 steps
HunyuanVideo / video	322.26	720 x 1,280; 129 frames at 15 fps; 50 steps

All rows: one B200 GPU. Video batch settings are 16, 8, 16 and 1 respectively. Clip lengths, resolution and frame counts differ, so these are examples rather than a direct efficiency ranking. Native sequence conventions can include an extra frame; frame counts are provided instead of claiming precisely equal clip lengths.[33]

Count candidates, not just final assets

Suppose a design needs 10 accepted images and generates four candidates per image. At an assumed full-facility 3 Wh per candidate, generation uses 10 x 4 x 3 = 120 Wh. Editing at a 60 W workstation for two hours adds another 120 Wh, before network and storage differences. The 3 Wh rate is a planning input, not a measurement of GPT-Image or any other named service.

For a roughly one-minute video assembled from eight short shots, four candidates per shot means 32 generations. Assumed facility rates of 30, 150 and 600 Wh per candidate give 0.96, 4.8 and 19.2 kWh respectively. Editing, audio, rendering and delivery are additional. These are sensitivity cases, not statistical bounds.

Making a single clip longer or increasing resolution may raise energy faster than proportionally. Do not extrapolate a short-clip benchmark to a long continuous video solely by multiplying seconds. Compare against the actual alternative production workflow, including travel and equipment only when they would really be avoided.

Source notes

33. ML.ENERGY research team. Leaderboard task datasets

Latest proprietary video estimates

A July 2026 preprint estimates energy for commercial video models from observed API timings and assumed hardware. Its Figure 1 reports the following for an eight-second 720p generation. These are inferred accelerator-level values, not provider meter readings or verified full-facility averages.[35]

Tested model	Estimated Wh / clip	Evidence status
Veo 3 Fast / Veo 3.1 Fast	27.0 / 27.0	API timing plus inferred hardware
Veo 3 / Veo 3.1	30.8 / 39.5	API timing plus inferred hardware
Seedance 1 / Seedance 1.5	80.0 / 102.1	API timing plus inferred hardware
Runway Gen 4.5	322.0	API timing plus inferred hardware
Sora 2 Pro	418.5	Tested version; not a claim of current availability

The same figure reports measured open-model values around 383-479 Wh on an eight-H200 setup. This does not conflict with a lower one-B200 benchmark: hardware, frames, steps and configuration differ. The commercial estimates assume deployments the researchers cannot inspect. Queueing, sharing and actual hardware can move the answer materially; statistical uncertainty does not capture every wrong assumption.[35]

Use these figures to understand possible scale, not to announce a precise environmental winner. Native audio also complicates equal-output comparison. Newer versions require new measurement; a current product name alone is insufficient.

Practical decision rule

Budget the number of candidate clips before production. Use short previews to select a direction, then increase quality only for accepted shots. If a tool reports actual GPU-seconds or measured energy, retain that trace alongside duration, resolution and retries. If it reports only dollars or elapsed queue time, label the energy conversion as uncertain.

A single 100 Wh generation has the same electricity magnitude as about 4.5 example phone charges or a 10 W lamp running for ten hours. That comparison says nothing by itself about water consumption, carbon intensity, artistic value or the footprint of the production it replaces.

Source notes

35. Nidhal Jegham, Boris Gamazaychikov and Sasha Luccioni. Lights, Camera, Carbon: Architectural Scaling Laws for Video Generation Energy Consumption

7. Google. Lithium battery test summary

Estimating tokens without false precision

For a logged text workflow, use separate rates for uncached input, cached input and generated tokens, including hidden reasoning where disclosed. Repeatedly sending a long document can process far more input than its original length. Cached input is cheaper to process in many systems, but it is not automatically energy-free.

Facility electricity in Wh = uncached input tokens / 1,000 x input rate + cached input tokens / 1,000 x cache rate + generated tokens / 1,000 x generation rate + tool electricity. Rates must state their hardware and infrastructure boundary. A rate that already includes facility overhead must not receive another PUE multiplier.

Illustrative rates, Wh / 1,000 tokens	Efficient	Middle	Intensive
Uncached input	0.01	0.10	1.00
Cached input	0.001	0.010	0.100
Generated, including reasoning	0.10	1.00	5.00
100k uncached + 400k cached + 30k generated	4.4 Wh	44 Wh	290 Wh
Add assumed tool electricity	1 Wh	5 Wh	20 Wh
Total research-workflow budget	5.4 Wh	49 Wh	310 Wh

These rates are deliberately broad stress-test assumptions. They are not fitted estimates, confidence intervals, current model averages or guaranteed limits. They make the sensitivity visible when telemetry is missing. Actual workloads can fall outside them. Use measurements from the intended deployment to replace the inputs.

Why “one million tokens” is not enough information

At the middle assumed rates, one million uncached input tokens cost 100 Wh; one million cached input tokens cost 10 Wh; one million generated tokens cost 1,000 Wh. These are arithmetic examples, not empirical universal conversion factors. At 1 Wh per 1,000 generated tokens, the 22.2 Wh phone charge equals about 22,200 generated tokens; the 120 Wh porch light equals 120,000. Neither comparison includes the input needed to produce those tokens.

For self-hosted systems, measure total node power over time, subtract or allocate background consumption according to the question, divide shared energy across completed requests, and add facility overhead. TDP is a component’s design rating, not actual whole-system power. API response time multiplied by an assumed dedicated GPU count can be wrong when workloads share hardware.

Tools such as EcoLogits can make consistent model-based estimates and disclose methodology. Such estimates should remain distinguishable from direct metering. A useful procurement requirement is auditable energy per completed workload with workload, quality and boundary disclosed.[47]

Source notes

47. EcoLogits. LLM inference methodology

The right comparison for productivity

Three questions produce different answers. First, how much energy is allocated to one task? Second, how much electricity actually changes over the same period? Third, what is the lifecycle impact of delivering the same accepted result? Keep these separate: each is useful, but only the second directly estimates a short-run electricity saving.

Accounting choice	What is counted	Interpretation
Task allocation	Assign hourly workstation and building shares to hours spent on the task.	Useful for reporting energy per work unit; shorter tasks receive less overhead allocation.
Same-period counterfactual	Measure the office or home across the same four hours, including after AI finishes.	Tests whether the meter actually falls. Continued HVAC, lighting and background appliances remain.
Lifecycle comparison	Operational electricity plus hardware, buildings, training and other relevant impacts.	Requires comparable amortization on both sides; not all lifecycle impacts can be represented in kWh.

For both workflows, define the deliverable and acceptance test first. Include human preparation, verification, correction and handoff in AI-assisted time. If AI takes twenty minutes to draft but three hours to repair, the relevant human time is three hours and twenty minutes. If the result fails, include fallback work.

A practical equation

Net electricity saved = avoided workstation electricity + actually avoided lighting/HVAC/other electricity + avoided conventional cloud work - added AI model/tool electricity - any additional downstream electricity. Measure each term over the same period. Existing cloud services common to both workflows can cancel in the difference, but have not become zero-energy services.

For a 60 W workstation that sleeps at 3 W, reducing active work from 4 hours to 20 minutes avoids (60 - 3) x (4 - 1/3) = 209 Wh. If the building load does not change, 209 Wh is the break-even added AI budget. A 30 Wh job saves 179 Wh; a 300 Wh job adds 91 Wh.

The scenario treats the AI service’s added electricity as a fully allocated facility budget while holding unchanged building loads fixed. That is transparent but is not a pure short-run marginal grid model on both sides. A rigorous consequential study would also estimate how extra AI demand changes server capacity, utilization and electricity supply over the chosen horizon.

Exclude human food and basal metabolism from this operational-electricity comparison. People continue to live and eat after finishing work. Include commuting only for trips actually eliminated, not merely because one task became faster. Gas heating should be reported separately as fuel use, with carbon conversion where needed.

Office: four hours versus twenty minutes

All numbers on this page are illustrative engineering assumptions. Consider a six-person office. Its average HVAC electricity is 1.5 kW during the four-hour window, lighting 120 W and networking 30 W. Allocate these equally: 250 W HVAC, 20 W lighting and 5 W network per seat. A workstation averages 60 W active and 3 W asleep. Coffee receives the same 50 Wh per person in both workflows.

Manual task electricity allocated per seat = (0.060 + 0.020 + 0.250 + 0.005) kW x 4 h + 0.050 kWh coffee = 1.390 kWh. AI-assisted human time is 1/3 hour including review, with 30 Wh added remote model/tool electricity. These are assumptions, not measured office norms.[12,44,45]

Scenario	Manual kWh	AI kWh	Change / meaning
Allocate only active task time	1.390	0.192	86.2% less allocated energy; does not prove the office meter falls this much.
Same 4 hours; HVAC, lights and WiFi stay on; workstation sleeps	1.390	1.211	12.9% less electricity for the assigned seat share.
As above; personal 20 W lighting also turns off	1.390	1.138	18.2% less; only valid if the light can actually be switched off.
Same building; AI budget rises to 300 Wh	1.390	1.481	6.5% more despite the same time saving.
Person uses the time saved for more computer work	1.390	1.420	2.2% more; output may increase, but electricity does not fall.

The fixed-building AI calculation is 0.275 kW x 4 h + 0.060 kW x 1/3 h + 0.003 kW x 11/3 h + 0.050 kWh coffee + 0.030 kWh AI = 1.211 kWh. The lights, cooling and network are explicitly still running for all four hours.

If only one of six people finishes early, whole-office electricity falls from 8.340 to 8.161 kWh: just 2.1%. If all six finish early but shared systems remain on, it falls to 7.266 kWh: 12.9%. Do not multiply one person's share of HVAC savings by six unless the underlying HVAC load actually changes.

An external monitor near 25 W is supported by a historical US stock estimate; the combined 60 W workstation remains an assumption. Nameplate charger wattage is not measured average draw. Coffee brewing and warming are event and standby loads, so coffee is held constant instead of treating every saved work hour as avoided brewing.[12,45]

Source notes

12. US Department of Energy. Purchasing Energy Efficient Commercial Coffee Brewers

44. US Department of Energy. Purchasing Energy-Efficient Computers

45. US Department of Energy; Bryan Urban and colleagues. Energy Consumption of Displays in the United States

When the office actually closes early

The same six-person office can save much more if all occupants leave and building controls respond. Assume everyone finishes in twenty minutes, lights turn off, the network stays at 30 W, workstations sleep at 3 W, and average HVAC demand falls from 1.5 kW to 0.5 kW for the remaining 3 hours 40 minutes. The HVAC response is an explicit assumption.

Four-hour whole-office total	Manual kWh	AI kWh
HVAC	6.000	2.333
Lighting	0.480	0.040
Networking	0.120	0.120
Six workstations	1.440	0.186
Coffee, same six allocations	0.300	0.300
Added AI, 30 Wh per worker	0.000	0.180
Total	8.340	3.159

The resulting electricity reduction is 5.181 kWh, or 62.1%. It comes from an actual operating change, especially cooling. It is not the consequence of dividing a constant daily electricity bill among fewer work hours.

Cooling and heating require a longer check

Thermal mass, humidity control, minimum ventilation and central schedules can prevent an immediate demand reduction. After a thermostat setback, some energy is used to restore comfort. Compare the whole day or next occupied period, under similar weather, to include this recovery. Closing one room may not affect a central air handler.

In summer, people and computers add heat. As an illustrative upper calculation, removing 100 W of occupant heat plus 57 W of workstation heat for 3 hours 40 minutes avoids 0.576 kWh of heat. At an assumed cooling coefficient of performance of 3, that corresponds to 0.192 kWh of cooling electricity, only if cooling actually responds. Do not add this saving on top of a measured HVAC reduction that already includes it.

In winter, reducing internal heat gains can increase heating demand. A gas furnace uses fuel plus electricity for fans; electric resistance heating and heat pumps behave differently. The net greenhouse impact depends on the heating system and both local and remote power sources.

The workstation-only break-even budget is 209 Wh per worker. If personal lighting also turns off it becomes about 282 Wh. For the whole-office closure case, the threshold is far higher because real HVAC and lighting electricity are avoided. The calculator exposes these terms separately.

Ontario homes: a specific summer model

Outdoor temperature alone cannot determine AC electricity. At 25°C outdoors, a shaded well-insulated home with a 25°C setpoint may need little cooling. A sunny house with a 22°C setpoint, substantial humidity and internal heat may still need it. Floor area, air leakage, roof and window exposure, equipment efficiency and the full day's weather matter.[9,10]

Illustrative home: a 185 m² (about 2,000 ft²) detached house, gas space heating and gas water heating, no pool or EV charging. Assume 400 W average background electricity including refrigerator and router (9.6 kWh/day), plus 2.4 kWh/day of cooking, laundry and other events. Add a 60 W workstation and 10 W task lamp for eight hours (0.56 kWh). Central AC draws 3 kW including blower when fully operating.

Daily cooling assumption	AC kWh/day	Whole home kWh/day	How to interpret it
No cooling	0	12.56	Possible on a mild day; still uses background and activity electricity.
2 full-load-equivalent hours	6	18.56	Modest cooling example; possible at 25°C, not a weather-derived forecast.
5 equivalent hours	15	27.56	More cooling, solar or humidity load, or lower setpoint.
8 equivalent hours	24	36.56	Stress case, not a claimed typical 25°C day.

An equivalent full-load hour is a unit of accumulated operation, not necessarily one continuous hour. A variable-speed system drawing 1.5 kW for four hours uses the same 6 kWh as a 3 kW system operating for two. Runtime and power must be measured or estimated together.

What actual Ontario evidence can establish

IESO's December 2023 whitepaper reports postal-area summer monthly residential averages of 798 kWh in 2019, 938 in 2020 and 895 in 2021 in its analysed data. Roughly dividing by 30 gives 27-31 kWh/day. These are historical geographic averages, not a current household-weighted detached-home average and not a prediction for a 25°C day.[42]

Toronto Hydro's September 2025 residential peak-demand figures help size electrical service. Peak kW is not a home's average draw for 24 hours; multiplying a peak by every hour would badly overstate consumption.[43]

Use twelve months of your own bills to establish seasonality, hourly smart-meter data to match weather and occupancy, and a plug meter for workstation power. A summer bill divided by days provides a daily total, but assigning all of it to work does not make it avoidable.

Source notes

9. Natural Resources Canada. Air Conditioning Your Home
10. Natural Resources Canada. Keeping the Heat In Section 9
42. IESO. Demand and Conservation Planning Research Whitepaper
43. Toronto Hydro. Typical residential peak demand values

Home: what changes after AI finishes

Compare the same four afternoon hours. Keep background load at 400 W and coffee at 50 Wh. Manual work uses a 60 W workstation plus 10 W task lamp for four hours. AI uses both for twenty minutes, then the workstation sleeps at 3 W and the lamp switches off; the remote AI budget is 30 Wh. Everyone stays home and the AC schedule remains unchanged.

AC load during 4-hour window	Manual kWh	AI kWh	Electricity reduction
No AC	1.930	1.714	0.216 kWh / 11.2%
3 kW x 25% duty = 0.75 kW average	4.930	4.714	0.216 kWh / 4.4%
3 kW x 60% duty = 1.80 kW average	9.130	8.914	0.216 kWh / 2.4%

The device and lamp avoid (60 - 3 + 10) W x 3 hours 40 minutes = 245.7 Wh. Subtracting 30 Wh of remote AI gives 215.7 Wh saved. Changing the fixed background or cooling load changes the percentage, but not the absolute saving. The refrigerator and router still run.

If the entire household leaves

In the moderate-cooling case, suppose AC averages 0.75 kW for the first twenty minutes and 0.30 kW for the remaining time after a setback. Four-hour AC energy becomes 1.35 rather than 3.00 kWh. The total AI-assisted case becomes 3.064 kWh versus 4.930 kWh, a 37.8% reduction. This assumes a real thermostat response and still requires a whole-day recovery check.

If one worker stops but other occupants need the same comfort, do not claim that setback saving. If the worker continues using the computer for another task, entertainment or more AI work, the avoided-device term may also disappear. What happens with the time matters more than assigning the house’s full daily bill to a task.

Multiple workers and multiple homes

Two people in one home share a single background load and HVAC system. Count the house once, then add the two workstations and AI workflows. Two people in separate homes need two actual household profiles. Closing a central office while everyone works from home can shift heating and cooling demand to several houses; compare buildings over the same period, including whether the office truly changes operation.

For a home with electric hot water, an EV, pool pumps or electric resistance heating, replace the example load profile. These can dominate the bill yet remain unrelated to whether a proposal took twenty minutes or four hours. Do not infer a causal energy saving from a lower allocation of those loads.

Large proposals: calendar time and work time

A project that lasts three months does not necessarily involve three months of continuous labour from every team member. Define person-hours. Example: five people x twelve weeks x forty hours/week x 25% of time on a proposal = 600 person-hours. This baseline is an illustrative project scope, not an observed industry average.

Assume an AI-assisted accepted proposal still needs 180 human hours: 60 for requirements and inputs, 40 for research and evidence, 20 for drafting and 60 for checking pricing, commitments, compliance and final approval. Budget 5 kWh for all added AI and associated tools, net of cloud work common to both approaches. This budget must be replaced with logs for a real project.

600-hour proposal example	Manual kWh	AI kWh	What the result means
Allocate 335 W per active person-hour	201.00	65.30	67.5% less allocated work energy.
Same building operations; only active device hours fall	201.00	182.06	9.4% less actual scenario electricity.
Same operations; AI budget 50 instead of 5 kWh	201.00	227.06	13.0% more electricity despite fewer human hours.

The fixed-building calculation retains 275 W of overhead across all 600 baseline person-hours: 165 kWh. AI-era devices use 60 W x 180 hours = 10.8 kWh active plus 3 W x 420 displaced hours = 1.26 kWh asleep. Add 5 kWh AI to get 182.06 kWh. This assumes desks otherwise would have been active during those hours, and the building still operates.

Device-only break-even AI electricity is 420 hours x (60 - 3) W = 23.94 kWh. If reviewers spend longer than 180 hours, or the AI run consumes more than 23.94 kWh, this fixed-building example may stop saving electricity. If office hours or space can actually shrink, add measured reductions separately.

A draft in minutes is a different deliverable

Generating headings, persuasive prose or a first proposal draft can be quick. Obtaining valid evidence, accurate pricing, legal review and an authorized commitment may remain the bottleneck. Compare accepted proposals of the same scope. If the intended deliverable really is only a draft, compare draft-production effort on both sides.

Coffee, commuting, cloud services and other event loads are omitted from these project totals when assumed identical and therefore cancelling. The kWh totals are the defined device-plus-building-share boundary, not a complete corporate footprint. Separate homes can be substituted using the home methodology.

Websites and longer software projects

Illustrative website build: four people x twelve weeks x forty hours/week x 50% effort = 960 person-hours. Suppose an AI-assisted workflow requires 360 human hours for requirements, content, design decisions, integration, testing, accessibility, security and deployment. Assume 20 kWh of added model/tool electricity above common conventional cloud work.

960-hour website example	Manual kWh	AI kWh	Change
Allocate 335 W per active person-hour	321.60	140.60	56.3% less allocated energy
Same building; devices sleep during displaced hours	321.60	307.40	4.4% less scenario electricity
Same building; added AI/tool budget 100 kWh	321.60	387.40	20.5% more scenario electricity

The fixed-building case retains 264 kWh overhead. Devices use 21.6 kWh active plus 1.8 kWh asleep during 600 displaced hours, then add 20 kWh of AI. The device-only break-even budget is 600 x 57 W = 34.2 kWh. All project labour and AI budgets here are assumptions, not measured productivity findings.

What recent productivity evidence actually shows

Epoch's 2026 MirrorCode work includes an Opus 4.7 recreation of a roughly 16,000-line command-line program in about fourteen hours, with 2,000 of 2,001 tests passing. Estimated human effort was two to seventeen weeks. That is a substantial demonstration, but estimated human time is not a randomized matched comparison; the output still has scope and validation limits.[46]

METR's February 2026 update cautions that later developer speedup estimates are affected by selection issues. Its earlier slowdown finding is also not a timeless verdict about newer tools. Neither supports a universal claim that software projects become hundreds of times faster.[14]

A simple template-based site can be a different case. Assume eight manual hours versus two AI-assisted human hours, with 0.1 kWh added AI. At a 57 W active-to-sleep difference, device energy saved is 6 x 0.057 - 0.1 = 0.242 kWh. The same method works without assuming every enterprise project is comparable to a landing page.

Count the website after launch

A persistent extra 1 W of attributable hosting demand is 8.76 kWh per year; 10 W is 87.6 kWh. This is arithmetic, not a hosting benchmark. Match traffic, functionality and performance. AI-created code that requires more compute or larger downloads may offset development savings; well-optimized code may reduce them. Include ongoing tests, builds and maintenance on both sides.

Source notes

14. METR. We are Changing our Developer Productivity Experiment Design

46. Epoch AI. MirrorCode

Water and carbon are separate calculations

Electricity equivalence is not carbon equivalence. For carbon, multiply each location's electricity by an appropriate grid emissions factor, keeping operational emissions separate from embodied emissions. Average grid factors support attribution; marginal factors may better describe the effect of a change in demand. Match time, geography and accounting method.

Illustrative carbon counterexample: an office saves 0.209 kWh locally but adds 0.030 kWh of remote AI. At assumed factors of 50 g CO2e/kWh locally and 400 g remotely, avoided emissions are 10.45 g and added emissions are 12 g. Electricity falls by 0.179 kWh, yet modeled operational emissions rise by 1.55 g. These factors are hypothetical; neither is asserted as the measured current Ontario or provider factor.

Ontario's generation mix is relatively low-carbon but not fossil-free. IESO reports 48.2% nuclear, 23.1% hydro and 19.3% gas for transmission-connected generation in 2025, alongside other sources. This is not itself a consumption-based emissions factor and excludes embedded generation. Imports and the time of demand matter.[15]

Direct and indirect water

Direct water consumption can occur when cooling water evaporates. Indirect water consumption occurs in electricity generation and supply chains. Withdrawal is water taken from a source; consumption is the share not promptly returned in comparable form. Do not add these as if they were the same measurement. Potable water, reclaimed water and water consumed in a stressed basin have different implications.[16]

If facility electricity is E kWh and PUE is 1.2, IT electricity is E / 1.2. With an assumed on-site WUE of 0.5 L per IT-kWh and electricity-supply consumption of 1 L per facility-kWh, water = (E / 1.2) x 0.5 + E x 1. A 0.030 kWh AI job gives 0.0425 L, or 42.5 mL; a 3 kWh job gives 4.25 L. These are sensitivity assumptions, not universal AI water rates.

Check the WUE denominator: some sources use different boundaries. If a water estimate already includes electricity generation, do not add it again. Compare water avoided from conventional office electricity too, using that location's factor; do not assume it equals the data centre's.

Google reports 0.26 mL of direct cooling water associated with its median May 2025 text request. Mistral's 2025 analysis reports a broader lifecycle inference estimate of 45 mL for a 400-token response. Different models and accounting boundaries prevent treating these as contradictory measurements of the same activity.[2,21]

Source notes

2. Google researchers. Measuring the environmental impact of delivering AI at Google Scale

15. IESO. Transmission Connected Resources

16. Arman Shehabi and colleagues, Lawrence Berkeley National Laboratory. 2024 United States Data Center Energy Usage Report

21. Mistral AI with Carbone 4 and ADEME. Our contribution to a global environmental standard for AI

The aggregate environmental picture

Small individual requests and large system demand can coexist. IEA's April 2026 assessment estimates all data centres consumed 485 TWh in 2025 and projects approximately 950 TWh in 2030, about 3% of global electricity. AI-focused data centres are a subset, projected at around 465 TWh in 2030. These categories include more than consumer text inference.[1]

For scale, one billion requests at an assumed arithmetic mean of 0.3 Wh would use 0.3 GWh per day, or 0.1095 TWh per year. At 30 Wh per job the same count would use 10.95 TWh per year. These are multiplication examples, not a forecast of actual request volumes or a legitimate use of a measured median as a mean.

LBNL estimates US data centres used 176 TWh in 2023, consumed about 66 billion litres of water directly and nearly 800 billion litres indirectly through electricity. These are all-data-centre estimates, not AI-only totals or all potable water. National totals can conceal severe local constraints where demand concentrates.[16]

Training and construction also count

Inference figures usually exclude model training and experimentation. Training is a large, intermittent cost; inference accumulates with use. A lifecycle analysis allocates training, chip manufacturing, data-centre construction and equipment replacement over the services they deliver. A simple division of training electricity by expected lifetime requests is an allocation assumption, sensitive to how widely the model is used.[18,19]

Apply a comparable lifecycle boundary to conventional work: laptops, office space and other equipment also have embodied impacts. Avoid claiming all of a worker's office or computer manufacturing is eliminated by shortening one task. Conversely, repeated large model launches and capacity expansion are not captured by a tiny per-query operating figure.

Efficiency and total demand can move in opposite directions

Google's 2026 environmental reporting says data-centre electricity rose 37% in 2025. More efficient chips, models and cooling can reduce electricity per result while total use grows with adoption and more ambitious workloads. Productivity improvements can also encourage more output; that may be economically valuable without being an absolute environmental saving.[27]

Local grid capacity, new generation, transmission, backup generators, construction and water stress need evaluation alongside global percentages. Annual clean-energy purchases do not guarantee a facility consumes carbon-free electricity every hour, and a global water-replenishment project does not necessarily repair a local seasonal shortage.

Source notes

1. IEA. Key Questions on Energy and AI

16. Arman Shehabi and colleagues, Lawrence Berkeley National Laboratory. 2024 United States Data Center Energy Usage Report

18. Alexandra Sasha Luccioni, Sylvain Viguier and Anne-Laure Ligozat. Estimating the Carbon Footprint of BLOOM a 176B Parameter Language Model

19. Meta. Llama 3.1 Model Card

27. Kate Brandt, Google. Read our 11th annual Environmental Report

Solutions and what their claims mean

Approach	Potential benefit	What to verify
Smaller models, routing, caching, batching, quantization	Less computation per acceptable result; improved server utilization.[34]	Quality, latency and total demand. Savings can be spent on more or longer jobs.
Direct liquid cooling and dry heat rejection	More effective heat removal; can reduce evaporative cooling demand.[23,24]	A liquid loop alone does not prove zero water consumption. Check the heat-rejection system and power penalty.
Reclaimed water and basin-aware siting	Reduces pressure on drinking-water supply; avoids vulnerable locations.[23]	Actual consumptive use, drought conditions, seasonal supply and local permits.
Heat recovery	Can displace building or district heat where there is a customer.[25]	Delivered useful heat, heat-pump electricity, distance and year-round demand; planned capacity is not delivered saving.
Clean electricity, storage and flexible demand	Reduces emissions and can ease peak grid pressure.[30]	Additional supply, hourly matching, delivery location and emissions when clean power is unavailable.
Longer hardware life and material recovery	Reduces embodied impacts and waste.[29]	Manufacturing footprint versus operational efficiency; full lifecycle trade-offs.

Microsoft's June 2026 discussion describes closed-loop designs and alternative cooling-water supplies. A 2025 Nature lifecycle study compares cooling technologies across multiple environmental measures. These approaches are real engineering developments, but their effects depend on climate, design and deployment; they do not eliminate upstream water or equipment manufacturing.[23,24]

Espoo's 2026 account describes the Microsoft-Fortum development for district heat recovery. Heat reuse is strongest when it replaces heat that otherwise would have been produced and when demand coincides with availability. The example is a development project, not evidence that every data centre can usefully export all its heat.[25]

"Net zero" usually concerns net greenhouse emissions, not zero electricity. Distinguish annual renewable matching, actual hourly carbon-free supply, operational emissions and full supply-chain net zero. Residual-emissions removals require credible durability and verification. Water-positive and replenishment claims likewise need local, time-specific accounting.[27,28,29,30]

The practical goal is lower total environmental impact for useful services: efficient workloads, cleaner supply, responsible siting and transparent measurement. A facility can improve one metric while worsening another, such as using less cooling water but more electricity.

Source notes

23. Judy Priest, Microsoft. Inside Microsoft two decade push to cut water intensity while scaling for growth
24. Husam Alissa and colleagues. Using life cycle assessment to drive innovation for sustainable cool clouds
25. City of Espoo. The Hepokorpi Project Data center supporting Espoo Carbon Neutral Future
27. Kate Brandt, Google. Read our 11th annual Environmental Report
28. Brad Smith and Melanie Nakagawa, Microsoft. Responsibly building the AI future
29. Google. 2026 Environmental Report highlights
30. IEA. Energy supply for AI
34. Jae-Won Chung and colleagues. Where Do the Joules Go? Diagnosing Inference Energy Consumption

A measurement plan people can examine

Define the job and observe both workflows

Choose a repeatable task and a clear acceptance test. Record human hours, AI model/version, all model calls, token categories, retries, searches, code execution and generated candidates. Include checking and revisions. For proposals, acceptance includes valid evidence and approved commitments; for websites, it includes agreed functionality, accessibility, tests and operational readiness.

Measure workstation wall power during active work, idle and sleep. Use a plug meter rather than the charger label. Collect hourly building or household consumption and note occupancy, thermostat settings and weather. Compare equal periods and check the following day when HVAC recovery could matter. Where feasible, repeat matched tasks and report the distribution, not just the best run.

Publish four separate results

Result	Required disclosure
Electricity per accepted result	Manual and AI-assisted workflow energy, success rates, human review and retries; kWh with boundary.
Actual same-period electricity change	Which equipment stopped, which stayed on, and what happened during time saved.
Operational carbon and water	Location/time factors, direct versus indirect water, consumption versus withdrawal, uncertainty.
Lifecycle and scale implications	Training and hardware allocation, facilities, additional output and longer-run capacity changes.

The workbook's office, home and project pages are deterministic scenarios. They are not forecasts or measured averages. Positive "saved" values mean the AI-assisted scenario uses less electricity within the defined boundary; negative values mean more. Input assumptions are intentionally editable, and benchmark rows retain source dates and system boundaries.

What can be said in an everyday conversation

"A short AI answer can use a fraction of the electricity of charging a phone, but research agents and video are much heavier. If AI finishes work sooner, it saves electricity only where equipment or building demand actually falls. Data centres still have a substantial and growing environmental footprint, so efficient use and cleaner infrastructure both matter."

This framing neither assigns guilt to every question nor excuses unlimited generation. It ties the environmental claim to a useful result, a real alternative and a measurement people can inspect.

Sources

Original publications and documentation are linked below. Access date for all sources is 14 September 2026. Company disclosures are identified as such; engineering scenarios in this report are calculations rather than measured findings.

1. IEA. Key Questions on Energy and AI

16 April 2026; executive summary and pp. 25-26. Used for: 2025 electricity estimate, 2030 forecast and intensive AI workloads.

2. Google researchers. Measuring the environmental impact of delivering AI at Google Scale

August 2025; May 2025 measurements; sections 3-4. Used for: Median Gemini energy and water; measurement boundaries.

3. Josh You, Epoch AI. How much energy does ChatGPT use

7 February 2025. Used for: GPT-4o estimate, 500 output tokens and long-input sensitivity.

5. George Kamiya, IEA. The carbon footprint of streaming video fact checking the headlines

2020; updated 11 December 2020; 2019 reference year. Used for: 77 Wh per hour of Netflix streaming.

6. Carbon Trust. Carbon impact of video streaming

June 2021; 2020 European data; Table 4, PDF p. 52. Used for: 188 Wh per hour and allocation of network/router electricity.

7. Google. Lithium battery test summary

Live manufacturer documentation; Pixel 9 test dated 23 January 2024. Used for: Pixel 9 battery rated 17.77 Wh.

8. US Department of Energy. Purchasing Energy Efficient Light Bulbs

Live guidance; accessed 14 September 2026. Used for: LED and incandescent wattage comparison.

9. Natural Resources Canada. Air Conditioning Your Home

Undated legacy technical guide; accessed 14 September 2026. Used for: Cooling load determinants and EER definition; not current equipment standards.

10. Natural Resources Canada. Keeping the Heat In Section 9

Live guidance; accessed 14 September 2026. Used for: Thermostats, cooling, humidity and building controls.

Sources continued

11. Virginia Cooperative Extension. Energy Series Estimating Appliance and Home Electronic Energy Use

2020. Used for: Watts times hours calculation and appliance variation.

12. US Department of Energy. Purchasing Energy Efficient Commercial Coffee Brewers

Live guidance; accessed 14 September 2026. Used for: Brewing, standby and energy saving modes.

14. METR. We are Changing our Developer Productivity Experiment Design

24 February 2026. Used for: 2025 slowdown finding and limitations of later speedup evidence.

15. IESO. Transmission Connected Resources

2025 annual output table; accessed 14 September 2026. Used for: Ontario generation mix; excludes embedded generation.

16. Arman Shehabi and colleagues, Lawrence Berkeley National Laboratory. 2024 United States Data Center Energy Usage Report

December 2024; executive summary and water analysis, PDF pp. 55-57. Used for: US electricity and direct/indirect water estimates; original report hosted by MIT.

18. Alexandra Sasha Luccioni, Sylvain Viguier and Anne-Laure Ligozat. Estimating the Carbon Footprint of BLOOM a 176B Parameter Language Model

Journal of Machine Learning Research 24, June 2023; Table 1 and section 4. Used for: BLOOM training energy and broader carbon accounting.

19. Meta. Llama 3.1 Model Card

July 2024. Used for: 30.84 million GPU hours and 700 W rating for 405B training.

21. Mistral AI with Carbone 4 and ADEME. Our contribution to a global environmental standard for AI

22 July 2025. Used for: Le Chat 400-token lifecycle inference water and carbon estimates.

23. Judy Priest, Microsoft. Inside Microsoft two decade push to cut water intensity while scaling for growth

24 June 2026. Used for: Closed-loop cooling, dry heat rejection and alternative water supplies.

Sources continued

24. Husam Alissa and colleagues. Using life cycle assessment to drive innovation for sustainable cool clouds

Nature 641, 331-338, 2025. Used for: Cooling technology lifecycle comparisons.

25. City of Espoo. The Hepokorpi Project Data center supporting Espoo Carbon Neutral Future

5 June 2026. Used for: Municipal description of Microsoft and Fortum heat-reuse development.

27. Kate Brandt, Google. Read our 11th annual Environmental Report

30 June 2026; 2025 performance. Used for: Electricity growth, operational and supply-chain emissions, clean energy contracts.

28. Brad Smith and Melanie Nakagawa, Microsoft. Responsibly building the AI future

9 July 2026; fiscal 2025 performance. Used for: Emissions growth, REC accounting change and water replenishment.

29. Google. 2026 Environmental Report highlights

June 2026; 2025 performance. Used for: 78 percent water replenishment and material circularity.

30. IEA. Energy supply for AI

Energy and AI, 10 April 2025. Used for: Electricity supply options, fossil generation and clean energy deployment.

32. Felipe Oviedo and colleagues. Energy Use of AI Inference, Efficiency Pathways, and Test-Time Scaling

Joule, 2026; arXiv v2 revised 9 June 2026. Used for: Updated modeled short and long request distributions; not product telemetry.

33. ML.ENERGY research team. Leaderboard task datasets

Repository snapshot accessed 14 September 2026; commit 5d899ea5a1c8934c0eff691df2b731d20490bb2a. Used for: Minimum-energy B200 configurations selected separately per model and task; J divided by 3,600.

34. Jae-Won Chung and colleagues. Where Do the Joules Go? Diagnosing Inference Energy Consumption

29 January 2026; arXiv 2601.22076. Used for: GPU measurement methodology, batching and task dependence.

Sources continued

35. Nidhal Jegham, Boris Gamazaychikov and Sasha Luccioni. Lights, Camera, Carbon: Architectural Scaling Laws for Video Generation Energy Consumption

5 July 2026 preprint, v1; Figure 1 and Appendix E. Used for: Measured open video models and inferred proprietary video energy; uncertainty limitations.

36. Kim, Shin, Chung and Rhu. The Cost of Dynamic Reasoning: Demystifying the Latency and Energy Overhead of Agentic Workflows

HPCA 2026; author manuscript v2, Table III. Used for: A100 GPU energy for HotpotQA agents; not commercial research-agent averages.

37. Pham and colleagues, Brave Research. AgentStop: Terminating Local AI Agents Early to Save Energy in Consumer Devices

CAIS, 26-29 May 2026; Table 2. Used for: M1 Max local agent benchmark; failures and success-adjusted energy.

38. Anthropic engineering. How we built our multi-agent research system

13 June 2025. Used for: Token amplification in a particular research-agent implementation.

39. OpenAI. Models

Live official catalogue, accessed 14 September 2026. Used for: Current model names; catalogue is not an energy disclosure.

40. Anthropic. Release notes

Entries through September 2026; accessed 14 September 2026. Used for: Current Claude releases; no comparable energy telemetry established.

41. Google AI for Developers. Gemini API changelog

Entries through September 2026; accessed 14 September 2026. Used for: Current Gemini releases; older Gemini Apps measurement is not model-specific.

42. IESO. Demand and Conservation Planning Research Whitepaper

December 2023; Table 17, p. 31; 2018-November 2021 underlying data. Used for: Historical postal-area summer residential consumption; not a 2026 detached-home mean.

43. Toronto Hydro. Typical residential peak demand values

September 2025 data; accessed 14 September 2026. Used for: Peak capacity is different from continuous average consumption.

Sources continued

44. US Department of Energy. Purchasing Energy-Efficient Computers

Live guidance, accessed 14 September 2026. Used for: Active, idle and sleep state distinctions; actual workstation draw must be measured.

45. US Department of Energy; Bryan Urban and colleagues. Energy Consumption of Displays in the United States

2022 presentation; 2020 stock estimates. Used for: Historical 25 W active external-monitor context, not a laptop measurement.

46. Epoch AI. MirrorCode

2026 benchmark; June 2026 paper and live results accessed 14 September 2026. Used for: Long coding-task demonstrations; estimated human effort is not a matched productivity trial.

47. EcoLogits. LLM inference methodology

Live methodology accessed 14 September 2026. Used for: Model-based environmental tracking; estimates should not be described as wall-meter readings.