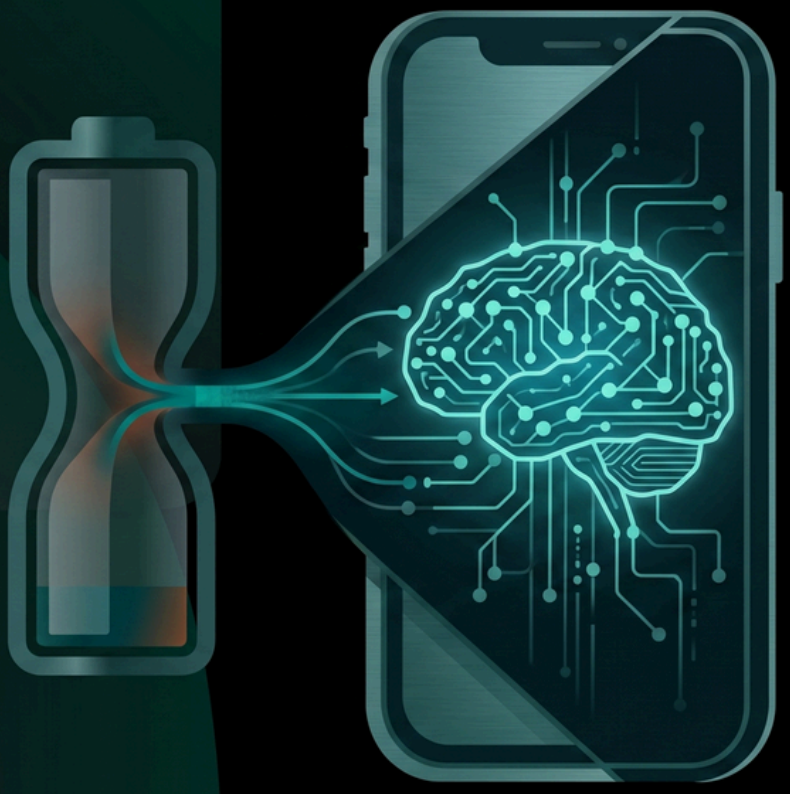
EDGE AI'S BATTERY BOTTLENECK:

ENERGY STORAGE LIMITATIONS FOR ON-DEVICE ARTIFICIAL INTELLIGENCE



Cornerstone Communications, ltd

Executive Summary

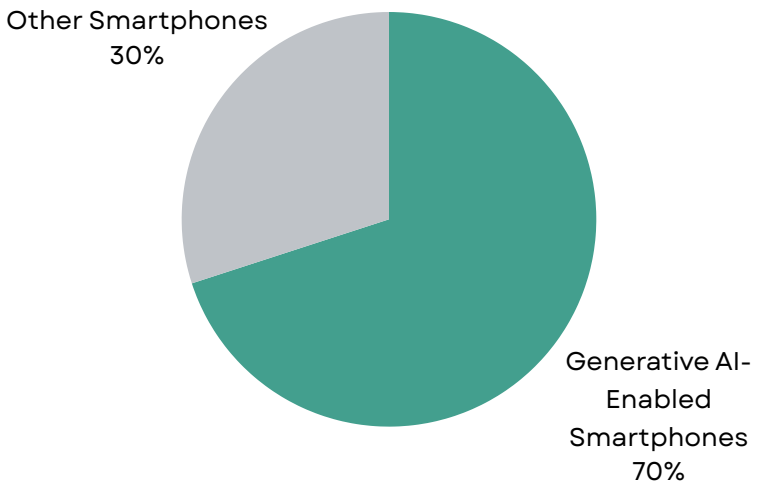
The global rollout of artificial intelligence (AI) faces a major barrier: the capacity of the electrical grid. The hyperscale data center model, which has been driven by the initial generative AI boom, is approaching a ceiling as data center energy demand is rapidly outpacing grid capacity.

Collectively, major technology companies, including Alphabet, Meta, Microsoft, and Amazon, allocated over $380 billion to AI infrastructure in 2025.[1] These investments will face severe power constraints. In major cloud computing hubs, local utility providers face massive backlogs. In 2024, Dominion Energy in Virginia had a data-center order book of 40 gigawatts, equivalent to the output of 40 nuclear reactors. That capacity has since risen to 47 gigawatts.[2] Globally, electricity consumption from data centers reached 415 terawatt-hours in 2024, representing about 1.5% of global electricity use. That is projected to more than double by 2030, with AI potentially driving 70% of that growth. Because building new high-voltage transmission lines requires a five-to-ten-year lead time, the industry is facing a potential energy crisis.

As data centers are being built at a rapid pace, the tech sector is exploring the potential to shift AI inference workloads to the "edge", i.e., moving calculations directly onto smartphones, laptops, and wearable devices. AI inference broadly refers to the process of running a trained AI model to make predictions. Executing these AI workloads locally instead of via the cloud can solve several key problems while improving the device user experience. On-device AI can decrease the latency of real-time applications like voice assistants and live translation, enable offline functionality when not connected to the internet, and improve data security and privacy by keeping personal data like photos and conversations solely on the device.

The consumer electronics market is moving rapidly towards this on-device AI model. The International Data Corporation (IDC) forecasts that shipments of generative AI-enabled smartphones will grow at a 78.4% CAGR, reaching 912 million units annually by 2028. At that stage, generative AI-enabled devices will represent over 70% of the entire global smartphone market.[3] Similarly, Canalys projected that AI-capable personal computers represented 40% of all PC shipments in 2025, climbing to 205 million units by 2028.[4]

Generative AI-enabled devices will represent over 70% of the entire global smartphone market in 2028



Relocating computation to consumers' devices shifts a massive computational and electrical burden from the grid-connected data centers to the devices in our hands. Chip makers and tech companies are actively attempting to address this problem by making the hardware and software work more efficiently for AI tasks. However, processing efficiency gains alone cannot keep up with the growing appetite for on-device AI. In parallel, the energy available to power the AI applications that are becoming increasingly integral to mobile devices must be increased. The energy density of conventional lithium-ion batteries has improved at a linear rate of approximately 5% per year.[5]

"As cloud-based AI is bottlenecked by grid capacity, a major limitation for the growth of edge AI is device battery life,"* states Dr. John Cooley, Founder and CEO of Nanoramic. Dr. Cooley is an energy storage expert who brought his battery technology company, Nanoramic, out of MIT and has over 15 years of experience commercializing advanced Li-ion energy storage. He adds, *"Without immediate leaps in energy storage technology, on-device AI faces the risk of being impractical for consumer adoption."

The Physics of On-Device AI Inference

To understand why AI drains batteries so rapidly, we must understand how a neural network operates compared to traditional mobile tasks and the energy consumption of those tasks. A typical premium smartphone contains a battery with a total capacity of approximately 13.5 to 15 Watt-hours (Wh). Traditional mobile computing tasks, such as streaming a high-definition video, operate via highly optimized, low-power processes utilizing dedicated hardware decoders. On an iPhone 16 Pro (which houses a roughly 13.6 Wh battery), streaming video draws an average of 0.4 to 0.6 Watts, allowing the device to achieve up to 27 hours of continuous playback.[6] Similarly, a 13-inch M3 MacBook Air draws a base power of approximately 5 Watts during light web surfing.[7]

AI inference (the process of running a trained AI model to make predictions) operates on an entirely different principle. Large Language Models generate output autoregressively, calculating one token at a time. For every single token generated, the device must move the entire multi-gigabyte neural network model from the system memory into the processor. Moving data across the motherboard requires orders of magnitude more energy than the mathematical calculation itself. The energy consumed by data movement between compute and memory is roughly 10,000 times larger than the energy used for the actual computation.[8]

Measurements of lightweight, on-device models report energy consumption ranging from 0.93 to 1.76 milliwatt-hours (mWh) per token generated.[9] Quantifying this in terms of AI text generation, OpenAI reports[10]:

1 token ≈ 4 characters

1 token ≈ ¾ of a word

100 tokens ≈ 75 words

1–2 sentences ≈ 30 tokens

1 paragraph ≈ 100 tokens

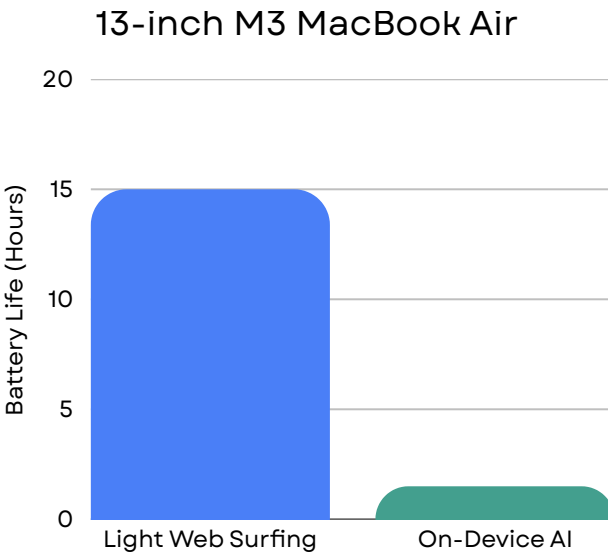
~1,500 words ≈ 2,048 tokens

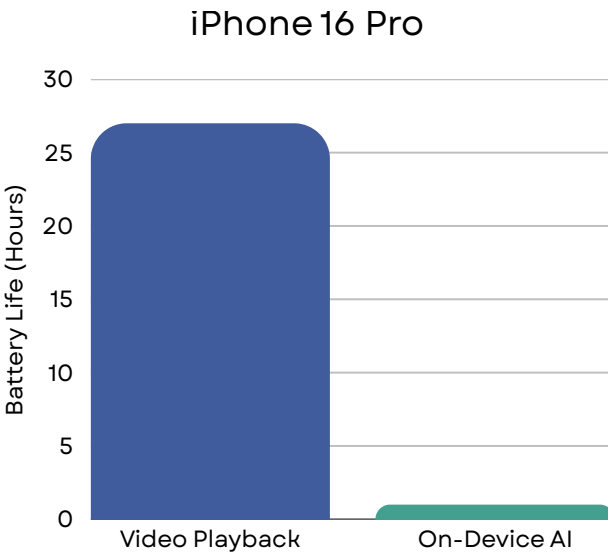
To put this into perspective, the energy consumption from 1000 tokens generated on-device can be up to 13% of the battery capacity of an iPhone 16 Pro.

A device sustaining maximum computational load is aggressively drawing power and generating significant heat. The iPhone 16 Pro draws up to 14.4 Watts under heavy load, and if sustained, this wattage would drain the entire battery in under an hour.[6] The M3 MacBook Air reaches a combined CPU and GPU peak of 33 Watts, a load that rapidly depletes its 52.6 Wh capacity.[7] As the demand for Edge AI increases, power consumption on devices will increase drastically and the need for improved thermal dissipation will become increasingly prevalent. This is a major point of tension that needs to be addressed. Without improvements to the batteries powering these devices, Edge AI faces a bottleneck towards its proliferation.

Estimated Battery Life for Consumer Device Workloads

Device	Workload Type	Sustained Power Draw	Estimated Battery Life
13-inch M3 MacBook Air	Light Web Surfing	~5.0 Watts	15+ Hours
	Heavy Compute for On-Device AI	33.0 Watts	~1.5 Hours
iPhone 16 Pro	Video Playback	~0.5 Watts	Up to 27 Hours
	Heavy Compute for On-Device AI	14.4 Watts	< 1 Hour





The Shift to Agentic Workloads

The power drain crisis is becoming increasingly prevalent due to AI software advancements. Early mobile AI features were reactive. A user pressed a button, the device processed a computational photography enhancement, and the system immediately returned to a low-power idle state. The current generation of software relies on Agentic AI. An AI agent operates autonomously in the background, continuously sensing contextual data, reasoning through multi-step logic paths, and executing tasks across different applications without direct user prompting. Agentic systems require continuous, sustained power draws rather than brief, burst power. Therefore, the processor remains under constant computational and thermal load.[11]

This thermal load and power drain has an immediate impact on user experience. Laptops equipped with advanced local AI copilots can experience up to a 50% reduction in battery runtime when these features are heavily utilized. Heavy mobile workloads can deplete a smartphone battery in as little as four hours.[5] Additionally, to prevent permanent chip damage and unsafe battery conditions from high heat, the system throttles the processor, degrading the accuracy and latency of the neural network.

Consumers are signaling frustration with this dynamic. An industry survey reveals that a smartphone's battery life remains the absolute highest priority for 53% of buyers, whereas AI capabilities rank fifth on their list of desired features.[12] Only 11% of consumers currently cite AI as their primary reason for upgrading a device.[12] If on-device AI requires sacrificing reliability and battery longevity, mainstream adoption will face severe headwinds.

53% of buyers say that the smartphone's battery life is the highest priority

11% of buyers say that AI is their primary reason for upgrading a device

Chip manufacturers are attempting to solve the computational energy problem by making processors more efficient. Modern mobile device chips designed for edge AI computing contain a Neural Processing Unit (NPU). Executing an AI workload on an NPU is exponentially more efficient than utilizing a standard CPU or GPU, which theoretically translates to improved battery life. Despite these optimizations, the continuous nature of Agentic AI forces the device to maintain high power draws for longer durations. In the near term, chip efficiency gains can reduce the amount of energy consumed per task. In the long term, battery improvements must be made to increase the total amount of energy available to spend. With smartphone sizes having reached their maximum, additional improvements in battery energy density are critical. As AI applications become increasingly integral to mobile devices, the demand for batteries that can support these power-intensive functions will continue to grow.

What the Energy Storage Industry Must Solve

To support the sustained thermal and electrical demands of on-device AI, the battery industry must accelerate past the incremental annual gains of conventional lithium-ion architectures.

A successful transition to the edge requires immediate technological advancements across these critical vectors:

Breaking the Energy Density Ceiling

Consumers are unwilling to accept thicker, heavier smartphones in exchange for AI computing performance and larger batteries. To power continuous background AI tasks within the existing physical footprint of consumer devices, the industry must adopt technologies that can significantly improve cell energy and power density.

Overview of Technological Advancements Used to Improve Battery Energy Density

Technological Advancement	Description
Higher mass loading	A conventional cell consists of copper and aluminum current collectors coated with active materials. The metallic foils themselves provide no energy storage. By applying thicker active material coatings onto these collectors, hence increasing the mass loading, manufacturers maximize the ratio of active materials to inactive current collectors. This directly improves the energy density of the cell.
Advanced cathode materials	The cathode active material plays a governing role in the battery cell's capacity. Manufacturers can maximize the energy density within existing footprints by transitioning to more advanced materials such as high-Nickel chemistries.
Advanced binders and additives	Standard polymer binders, such as Polyvinylidene Fluoride (PVDF), provide mechanical adhesion between the active materials and to the current collector. However, they introduce thermal and electrical resistance, occupying critical internal volume without contributing to energy and power. Replacing these insulating polymers with advanced binders and additives can reduce internal resistance, aid in thermal dissipation, and enable higher active material content which in turn improves energy density.
Silicon anode	Conventional anodes are made with graphite as the active material. Manufacturers are adopting Silicon for the anode as the material offers a theoretical specific capacity nearly an order of magnitude higher than graphite. However, the main problem with Silicon is physical swelling. Therefore, commercialization of silicon anodes requires advanced binders and technologies to mitigate swelling.
Lithium metal anodes	Conventional anodes use graphite or silicon to host lithium. A lithium-metal architecture plates metallic lithium directly onto the current collector, achieving the theoretical maximum for anode energy density. The main hurdle for Li-metal anodes is the greater risk of dendrite formation: the growth of microscopic lithium filaments during charging that can pierce the separator and induce short circuits and thermal runaway.
Advanced electrolytes	Conventional liquid electrolytes transport lithium ions between electrodes, but can degrade at high voltage and thermal loads generated by AI computation. Advancements in liquid electrolyte that improve electrochemical stability can enable cells that safely store more energy and operate at higher temperatures.

Drastic Reductions in Electrical Resistance

When an Agentic AI system executes a complex reasoning chain, it demands a rapid, high-power pulse from the battery. When a device draws 15 Watts continuously, it effectively subjects a standard 15 Wh smartphone battery to a 1C discharge rate. A battery that delivers its full capacity during a slow trickle discharge might only deliver a fraction of that energy under a heavy AI workload due to energy lost as waste heat and practical limitations of lower terminal voltages under heavy loads.

"High internal electrical resistance translates to poor capacity retention under load,"* explains Dr. Cooley. *"The energy storage industry must adopt electrode designs with low resistance that allow the battery to yield its full rated capacity even under continuous workloads like what you would experience with on-device AI."

Sustainable Materials and Lower Carbon Footprint

As on-device AI grows, the increasing number of lithium-ion batteries being produced and reaching end-of-life presents a sustainability challenge. It is crucial for the industry to adopt new technologies with the ability to reduce the environmental impact of battery materials, manufacturing, and end-of-life processes. Cell makers can transition to manufacturing processes that reduce the carbon footprint caused by the energy-intensive electrode drying process, while simultaneously replacing destructive, high-heat smelting with direct recycling methods that enable circular materials.

"Sustainability is necessary for manufacturers to support this massive scale of battery production,"* notes Dr. Cooley. *"Transitioning to cleaner, more efficient manufacturing processes can be achieved in parallel with lowering costs and will help stabilize the supply chain for the transition to edge AI."

Investor Realignment

The trend toward edge AI has the potential to trigger a massive realignment in institutional investment. During the initial wave of AI funding, capital flowed almost exclusively into large language model developers and the construction of hyperscale data centers. As the physical limitations of the cloud model become undeniable and the benefits of on-device AI become increasingly prevalent, investors should look towards the intersection of edge AI and battery technology. Whoever solves the energy storage bottleneck at the edge will capture the massive value generated by the next generation of mobile hardware.

Conclusion

The trajectory of the technology sector is moving rapidly towards localized, on-device AI. The software capabilities of Agentic AI are changing the baseline expectations for consumer and enterprise hardware. A major friction point in this transition is energy storage. The computation required to run an AI agent on-device results in massive sustained power draws. The deployment of localized AI cannot rely solely on the optimization of software or improvement of chip hardware. Energy storage must evolve in parallel. The energy storage industry must aggressively commercialize new technologies that solve the power delivery, capacity retention, and energy density bottlenecks. Consumers will continue to prioritize basic reliability over AI features on their mobile devices. Rising prices of mobile devices, partially driven by AI's impact on the memory supply chain, combined with a lack of immediate battery innovation may push customers away from newer devices and present major risks to the consumer electronics manufacturers. Without a paradigm shift in battery technology, the edge AI revolution will stall.

GET IN TOUCH

240.301.3600
info@cornerstonepr.net
www.cornerstonepr.net

Cornerstone Communications, ltd

References

[1] Byteiota, "Google's $4.75B Power Play: AI Hits the Energy Wall", 2025.

[2] ET Telecom, "The urgent power crisis in the AI revolution", 2025.

[3] International Data Corporation (IDC), "Worldwide Generative AI Smartphone Shipments Forecast to Reach 70% of the Market by 2028", 2024.

[4] Canalys, "AI-Capable PCs Forecast to Make Up 40% of Global PC Shipments in 2025", 2024.

[5] Enovix, "The AI Power Drain: Why Battery Limitations Threaten the Future of Mobile AI", 2024.

[6] Notebookcheck, "Apple iPhone 16 Pro Smartphone Review", 2024.

[7] Notebookcheck, "Apple MacBook Air 13 M3 Review", 2024.

[8] Applied Materials, "Applied Materials Hosts Lively Debate on AI Energy Efficiency", 2023.

[9] Nguyen, V., et al., 2025, “On-Device or Remote? On the Energy Efficiency of Fetching LLM-Generated Content.”, In 2025 IEEE/ACM 4th International Conference on AI Engineering – Software Engineering for AI (CAIN).

[10] OpenAI, “What are tokens and how to count them?”, 2026.

[11] Bandi, A., et al., “The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges.”, Future Internet, 2025.

[12] CNET / Android Headlines, "Users Still Care More About Battery Life Than AI, Survey Reveals", 2025.