GenAI's Impact on Computing Power:

Lessons from a Bullish Model of Global Demand

By Azeem Azhar, François Candelon, Riccarda Joas, Nathan Warren, and David Zuluaga Martínez

September 2024

Introduction

There has been much anxiety in the last couple of years over the future availability of affordable computing power, the hardware that underpins the modern economy's digital infrastructure. Such worries are understandable for anyone tracking the rapid advancements in generative artificial intelligence (GenAI). The time it takes to double compute demand to train such models is now faster than Moore's Law—and poised to accelerate further, as the tech giants continue to bet on scale as the driver of progress in artificial intelligence (AI). The joint Microsoft and OpenAI plan for a $100 billion supercomputer is one recent indication of this trend. Spending on computing power by businesses has climbed alongside the growing compute intensity of AI: 2023 was the year when Google for the first time spent more on computing than people.

While demand for AI workloads continues to rise, the supply for specialized hardware, especially Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs)—manufactured by a very small number of companies—appears unable to keep pace. A rapid increase in supply is unlikely for two reasons. On the one hand, this sector is hard to disrupt by outsiders, as there are significant barriers to entry due to intellectual property rights on the complex designs and architecture of GPUs and TPUs. Catching up to the incumbents' IP would require substantial upfront R&D investment. On the other hand, an expansion of existing manufacturing capacity by incumbents is no easy feat given the complexity of the supply chain and sheer capital intensity of high-end chip production. The cost of setting up a chip production facility is estimated to have increased threefold in the last six years, from $7 billion in 2017 to $20 billion in 2023, mainly driven by the sophistication of the required machinery.

As hyperscalers such as Google or Microsoft build ever-larger GenAI models requiring more and more specialized hardware, won't we face a severe scarcity of computing power? As we argue in this piece, however, once one clearly understands the different computational and hardware demands of model *training* and model *inference*, this scenario starts to seem less likely than might appear at first. In fact, even under bullish assumptions about the growth and intensity of GenAI demand in the coming years, it is far from obvious that we're headed toward a structural scarcity of computing power. To test this claim, we built a quantitative model that is moderate in its supply estimates but bullish in its demand assumptions and found that GenAI workloads would account for only ~34% of global data center-based AI computing supply by 2028. The rise of GenAI, and the computational requirements associated with it, is unlikely to "break" the decades-long regime of affordable, widely available computing power.

Of course, there may be other factors that could severely constrain computing power supply—notably the energy required to power the data centers where much of the world's AI workloads are and will continue to be processed. But the central contention of our analysis remains: The rapid adoption of GenAI will not *by itself* outpace the world's capacity to produce the required computing hardware.

The historic development of computing power

There have been no serious concerns about the supply of computing power over the last five decades because of two factors operating in tandem: Moore's law and large-scale digitization. Since approximately 1970, the number of transistors per chip has roughly doubled every two years, with process nodes shrinking during that period by four orders of magnitude—enabling far more computations per chip. At the same time, computing hardware has itself become omnipresent in the form of data centers, computers, phones, and other devices. We estimate that the growing computing power of hardware, together with the explosion in the number of physical devices, has made the total supply of computing grow by approximately 60% CAGR since the 1970s.[1]

As a result of these trends, the world has largely operated under a regime of reliably affordable computing power relative to the computational demands of existing technologies. One telling indicator of the degree to which businesses have not faced meaningful computing supply constraints is the sheer inefficiency of their code base. A 2018 report by Stripe estimated that "bad code" costs companies $85 billion annually. The prevalence of "bad code" is at least suggestive that computing power has been somewhat like oil in the early 20th century: valuable, but plentiful enough to be used without much regard for optimization.

The computational demands of GenAI

The key question many have raised is whether GenAI will lead to a sudden change in the decades-old regime of ample availability of affordable computing power. Can supply possibly keep up with generative models' need for computing?

To answer this question, one must first get clear on the different computing needs associated with GenAI. Broadly speaking, there are three types of GenAI compute workloads: model training, fine-tuning, and inference. The distinction between these is critical because model training requires specialized hardware, whereas fine-tuning and inference *can* be run on less specialized chips, even if more specialized hardware is optimal and hence preferred whenever available.

Developing a foundation model involves large-scale training, which is resource-intensive and expensive. Consequently, only a handful of businesses engage in foundation model training—mostly hyperscalers or dedicated businesses with close ties to tech giants such as OpenAI or Anthropic. The size, data-richness, and market position of the very few non-hyperscalers that have developed their own models—such as

[1] To estimate the increase in global computing power since 1971, we multiply the average FLOPs/IPs per device category in both 1971 and 2023 by the total number of devices for the respective years. The device types we focus on are mainframes and minicomputers for 1971 and PCs, mobile devices, and servers for 2023. (FLOP, a common measure of computing performance, stands for floating point operation, or arithmetic operation performed on floating-point numbers, such as addition, subtraction, multiplication, and division.)

Bloomberg—highlight how rare these exceptions are likely to remain. The reality is that most businesses are likely to become consumers rather than developers of GenAI foundation models, even if many do engage in the more modest training effort of fine-tuning existing foundation models. Now, compared to foundation model training, fine-tuning is a far less computationally intensive effort, typically requiring less than 10% of the cost of foundation model pre-training and often less than 0.1%. Furthermore, once a model is suitably fine-tuned, it will only occasionally require retraining.

By contrast with both foundation model training and fine-tuning, *inference* refers to the actual use (or prompting) of a GenAI model. For any given model, the odds are that the total computing power used for inference far exceeds that used for model training, as a single model is meant to support a large number of users over an extended period of time.[2] Now, the crucial fact is that the specialized hardware needed to train the largest models is *not* required for inference. Inference workloads don't even need the same sort of data center infrastructure: Whereas training requires minimal transmission latency, and hence the physical co-location of the hardware in a single data center, inference can be carried out in a distributed fashion, optimizing utilization across multiple data centers. In fact, inference will increasingly take place on the edge—on people's laptops and phones, and in embedded systems, as with the vision for Apple Intelligence and Meta's smaller-sized LLaMA-7B.

That inference will tend to account for the lion's share of total GenAI workloads and is less reliant than model training on specialized hardware, already suggests that fears about computing power scarcity may be overstated. The key question, then, is this: How much GenAI model inference will there be, and how computationally intensive will it be? It is difficult to form reliable estimates of future use patterns for any nascent, general-purpose technology with such strong signs of exponential adoption. So instead of attempting a projection, we have opted to model a scenario of bullish GenAI inference demand and moderate computing power supply over the next four years to test the soundness of prevailing anxieties about computing power availability.

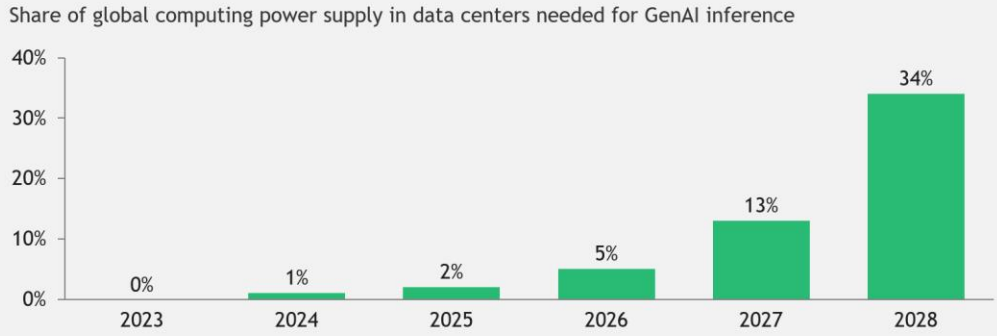
A bullish scenario of computing power demand driven by GenAI inference

The demand in our model is an aggregation of businesses, governments, and individual consumers using GenAI, with a focus on workloads that would require utilizing AI chips in data centers (as opposed to edge devices). The supply in our model is constituted by the availability of GPUs, setting aside the important fact that computations for model inference can run on less specialized hardware.

Our modeled aggregate global demand for GenAI model inference reaches ~34% of the total available data center computing power for AI by 2028 (Exhibit 1). Given our reliance on bullish demand assumptions and moderate supply projections—which we explain in detail in the following sections—this result strongly suggests that the rise of GenAI is unlikely to undermine the current regime of widely available, affordable computing power.

[2] Consider the case of Gemini Ultra, one of the frontier models requiring the most compute-intensive training to date: It would take 100 million users conducting a single medium-length interaction per day (e.g., drafting a short article with ~5e15 FLOP) for only 100 days to collectively utilize as much computing training for inference as was required to train the model.

Exhibit 1 | Demand for GenAI inference, even if bullish, would not exceed ~34% of computing power supply



Source: Exponential View and BCG Henderson Institute analysis.

Bullish demand

In our model, we structured the demand for computing power for GenAI inference around three assumptions: (1) the continuation of model scaling, leading to increasing computational intensity of GenAI training and inference over time; (2) an aggressive extensive margin of global GenAI adoption; and (3) a similarly aggressive intensive margin of adoption, with high GenAI utilization levels by businesses and consumers alike.

Demand Assumption #1: The scaling game continues

Our first assumption is that frontier GenAI models keep getting bigger, just as they have done for several years now, with the parameter count rising at 2.8x per year since 2018.[3] We do expect new generations of generative models to be larger—as anticipated by some of their leading developers—regardless of when (or whether) significant jumps in performance occur as a function of size. What makes our calculation bullish is the assumption that frontier models may get as large as engaging 15 trillion parameters per prompt by 2028.[4] This is more than 50 times as computationally intensive as the average

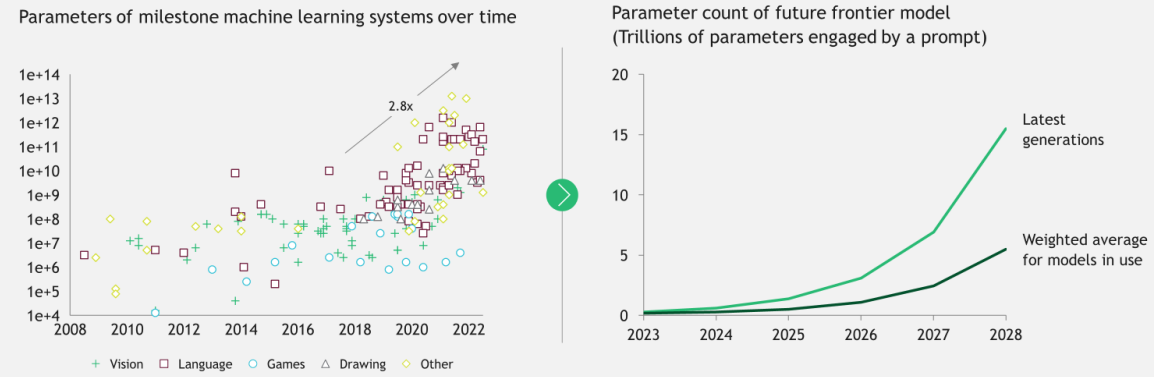
[3] In our model, a parameter growth of 2.236x per year is assumed. We derive this growth rate by incorporating a projected growth in total computing intensity of model training of ~5x per year according to estimates by EpochAI and assuming the Chinchilla scaling laws hold, which suggest that parameters and data, i.e., training tokens scale equally. Each therefore grows at the square root of 5 (2.236) per year. Our focus is only on parameter count because the compute intensity of *inference* depends solely on it (not on the volume of data used for model training).

[4] The cost of the training cluster for such a model would exceed $100 billion. While exorbitant, this value is in line with projects such as the Stargate data center.

GPT-4 inference.[5] Given the assumption that scaling will continue, the question is, "Which model will be most widely used?" It is unlikely that users will instantaneously upgrade their models every time a new, larger one is released. The cost of inference grows monotonically with model size, and when considering cost-to-benefit ratios, the latest frontier option won't be required for certain use cases. Further, businesses may face sizeable switching costs. We therefore assume that each generation of models remains in use for several years, with only "new" users directly turning to the latest—and largest—frontier options. In our model, by 2028, a user prompt will on average run through 5 trillion parameters.

For agentic workflows,[6] we assume smaller models will be used, where prompts would flow through approximately 3 trillion parameters—less than the average human prompt going through a frontier model, but still a very bullish assumption. In general, smaller, specialized alternatives are the more likely ones to be directly used by businesses and individuals over time. In particular, models built on domain-specific algorithmic and architectural innovations may be better suited in a variety of cases. Beyond the realm of internet text, specialized data sets are driving innovation across industries—from unraveling genetic codes to optimize drug development, to harnessing factory sensor data for predictive maintenance, and accelerating material discovery through computational chemistry.

Exhibit 2 | First bullish demand assumption: The scaling game continues



Source: EpochAI; Exponential View and BCG Henderson Institute analysis.

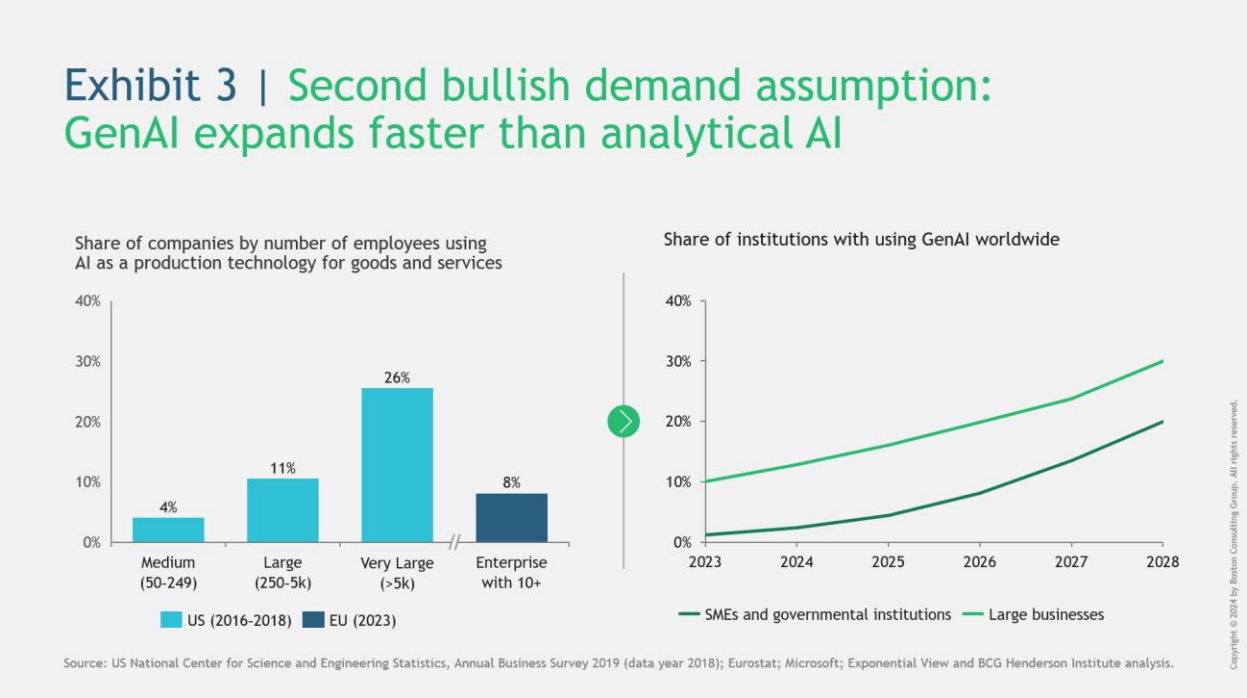
[5] For reference, GPT-4 is said to have 1.8 trillion parameters. However, GPT-4 is a "mixture of experts" model, not a monolith, rumored to consist of 16 discrete models, each with approximately 111 billion parameters, plus 55 billion shared attention parameters. The average prompt used on GPT-4 is estimated to engage two of those 16 models alongside the shared attention parameters, meaning that the prompt must "pass through" the outputs of approximately 277 billion parameters.

[6] By "agentic," we mean workflows that utilize AI systems specifically engineered to comprehend complex scenarios and autonomously achieve objectives, with minimal-to-no-human intervention.

Demand Assumption #2: Global GenAI adoption will far exceed the rate of analytical AI adoption

Setting aside the computational intensity of GenAI models over time, we imagine a "double exponential" adoption pattern for GenAI, in terms of the number of businesses, governments, and consumers using the technology (extensive margin) as well as the degree to which they use it (intensive margin).

As regards the extensive margin of institutional adoption, we assume that 20% of small and medium-sized enterprises (SMEs) and 30% of large businesses worldwide will be using GenAI substantially by 2028—as will 20% of all governmental institutions. Our assumptions about the global share of businesses using GenAI over the next few years are aggressive if one considers that, as late as 2018, the share of US businesses using AI for the production of goods and services was below 4% for SMEs, 11% for large companies with up to 5,000 employees, and only 26% among the very largest businesses (with over 5,000 employees). For the European Union as a whole, only 8% of businesses with more than ten employees used at least one AI technology as of 2021. The assumed 20% adoption rate for governmental institutions effectively treats these as comparable to large US corporations in terms of technological adoption—which is ambitious. Our institutional adoption assumptions are meant to reflect the genuine possibility that GenAI might be easier to deploy due to its natural language interface, but they remain aggressive in comparison with the observed rate of adoption of previous generations of AI, not least because we are making global extrapolations that exceed the pace of technological change in some of the world's most advanced economies.



There is one specific type of business application of GenAI that we have modeled independently, namely digital advertising. With increasing capabilities for multimodality—the ability to process inputs and generate outputs across data types, including images, audio, video, or sensory data—we expect digital advertising to become a natural use case for GenAI, if only because some of the world's leading GenAI

developers are also the largest players in digital advertising.[7] In keeping with the bullish spirit of our model, we assume that by 2028, all ads running on Meta's platforms—accounting for one-fifth of US digital ad spend—will display GenAI-powered, personalized images and captioning.

Lastly, we consider inference demand driven by individual consumers. It is considerably more difficult to anchor even bullish expectations for consumer GenAI demand—especially for the portion that will continue to require data center processing as opposed to edge computing. We therefore assume that the ratio of institutional-to-consumer demand in this space will resemble that of another large digital product with similar appeal to both segments: Microsoft Office. For our purposes, we assume that GenAI inference demand from individual consumers will be ~15% of that from institutions, about in line with Microsoft Office 365's reported ratio of enterprise-to-consumer revenue for 2022.[8] This 15% is meant to capture uses like chatbot interactions, features integrated into software such as Adobe, videoconferencing transcripts, and so on.

Demand Assumption #3: The intensity of GenAI use will grow exponentially

In addition to the rapid uptake of GenAI, our model assumes high and rapidly growing utilization—the intensive margin—among institutional and individual consumers alike.

In the case of SMEs, large businesses, and governments, we assume there will be two primary types of inference workloads. On the one hand, employees will individually use the technology in their day-to-day tasks. On the other hand, there will be more and more agentic workflows that automate end-to-end tasks or activities. For employees, we postulate individual worker utilization starting at an aggressive rate of 10,000 tokens per employee per day, roughly equivalent to about four short interactions with a GenAI user interface (e.g., asking a question of HR or summarizing an article) or one medium-length interaction (e.g., a brainstorming session or drafting a short article).[9] When it comes to agentic workflows, we make two assumptions. First, such workflows become widely available starting in 2025, which is sooner than many experts anticipate. Second—and this is the area of greatest uncertainty—we assume that an average agentic workflow will initially require about 500,000 tokens of inference per day in 2025, rising to about 2 million tokens by 2028. We arrived at this initial estimate by testing various approaches to end-to-end, automated research workflows.[10] As for the sheer volume of agentic workflows in any given

[7] We see further use cases for multimodal models beyond image and video generation, but do not expect them to be widely in use among businesses by 2028 given the substantial delay that typically occurs between technological breakthroughs and widespread commercial adoption.

[8] In keeping with the bullish tenor of our model, the implied computing power required by our consumer GenAI demand assumption is greater than that of all Google queries (based on available estimates of their total count and average token intensity).

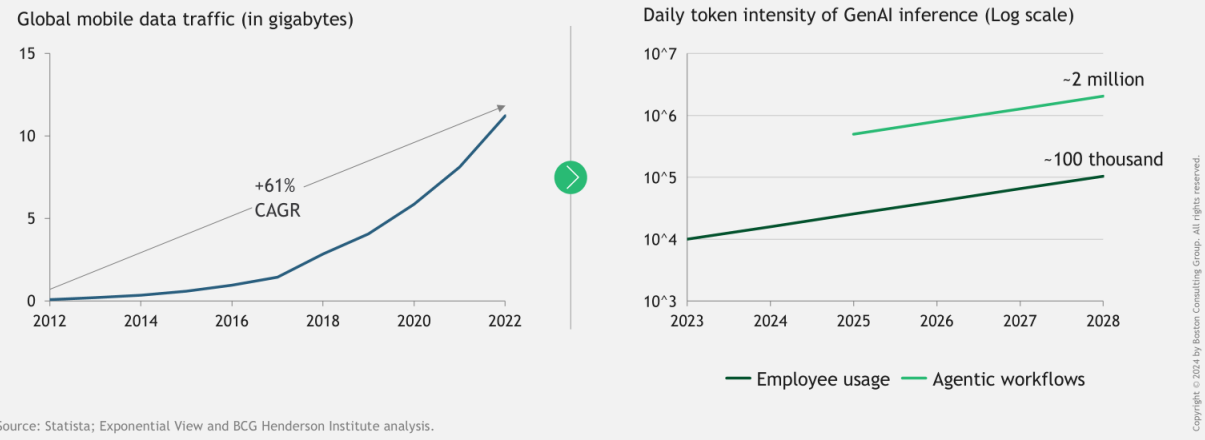
[9] Our token-intensity estimates by type of interaction are based on numerous test interactions with OpenAI's GPT-4 Turbo.

[10] As an illustration of what we mean by agentic workflows, consider a developer using GenAI for debugging. Assuming each developer works on an average of three problems per day, that each problem requires an average of three agentic debugging iterations, that the average problem size is 200 lines of code, and that each problem takes about 100,000 tokens to resolve, then daily token usage per developer add up to 300,000 tokens. Another example might be a researcher using GenAI. The AI researcher breaks down the research process into subtopics, generates

business, we assume that there will be one such workflow per employee per day in all businesses using GenAI, largely because we expect agentic workflows to still require some degree of human oversight.

Our baseline levels of token-intensity for both employee use and agentic workflows are aggressive—but not as much as the growth rate we apply to both. We assume that the intensity of use will grow at a rate comparable to mobile data traffic, one of the highest rates of growth in technology use in recent decades, estimated at 60% per year between 2012 and 2022.

Exhibit 4 | Third bullish demand assumption: The intensity of GenAI use will grow exponentially



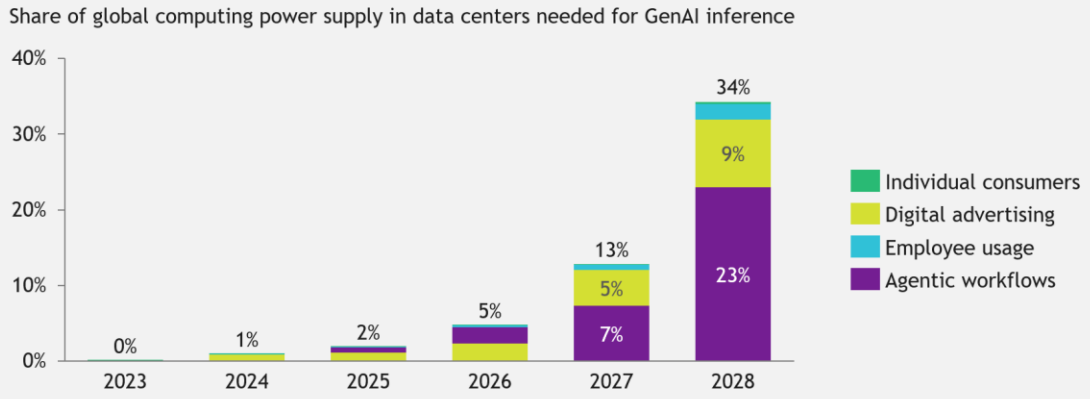
Source: Statista; Exponential View and BCG Henderson Institute analysis.

For demand due to digital advertising, we assume that Meta's ~3.3 billion daily users will encounter 30 personalized ads every day with automatically generated text and images. We keep these figures constant over time. As for individual consumers, recall that we fix the total inference demand at ~15% of total business demand. Since business demand expands over time—in both extensive and intensive margins—then, by implication, so does consumer demand as estimated in our model.

The full picture of our bullish demand scenario for data center–based GenAI inference is summarized in Exhibit 5. Most of the demand is driven by agentic workflows, which are at the upper end of what is feasible with today's technology. The crucial point is that even with our very aggressive assumptions, there will be sufficient supply, and GenAI inference—reaching ~2e30 FLOPs by 2028—would only take up ~34% of likely global supply.

individual reports for each subtopic, and combines them into a final comprehensive report; in our tests, this exercise took around 30 minutes and consumed between 1 and 2 million tokens per research brief.

Exhibit 5 | Agentic workflows are the biggest driver of GenAI inference demand in our bullish scenario



Source: Exponential View and BCG Henderson Institute analysis.

Moderate supply

We say that our supply figures are "likely" because, unlike the demand assumptions, here we adopt a moderate rather than bullish stance, in the spirit of testing the resilience of the current regime of widely available and affordable computing power.

First, we estimate today's baseline level of computing power available for GenAI inference as equivalent to the quantity of state-of-the-art (SOTA) GPUs already in use.[11] Virtually all those GPUs are at present produced by Nvidia. Recent reports indicate that Nvidia shipped ~3.8M data center GPUs in 2023, growing sales by ~42%—over a million more graphic processors—compared to 2022. In 2023, about 650,000 of these were at H100 chip-level performance and we assume the rest in 2023 as well as the aggregate in 2022 to be at A100 FP8 performance level. We arrive at a baseline quantity of available compute for GenAI workloads, assuming—plausibly—that all these chips are still in use at an average utilization rate of 50% (i.e., 50% utilization of their maximum theoretical FLOP per second due to limiting factors such as memory bandwidth). In this way, we arrive at a supply of ~7e28 FLOPs for the year 2023.

Having estimated supply in 2023, we then use SemiAnalysis's estimates suggesting that AI computing power overall will increase by roughly 60x by the end of 2025 compared to Q1 2023. We expect the growth rate (approximately 150% per year by Q4 2025) to soften, continuing to grow at ~60% per year through 2028. These growth rates might seem bullish, but they are in fact quite realistic. For starters, SemiAnalysis's view is premised on detailed data collection and analysis from the entire supply chain and

[11] As previously noted, this assumption sets aside the fact that computations for model inference can run on less specialized hardware—which *a fortiori* strengthens our argument.

major market players, not merely a "top-down" perspective. Furthermore, it takes account of supply chain limitations, particularly in critical components like Chip-on-Wafer-on-Substrate (CoWoS)[12] and High Bandwidth Memory (HBM).[13] In 2028, we expect a supply as high as ~4e30 FLOPs (or ~57 times as much as in 2023).

Supply may in fact turn out to be more plentiful than our figures suggest, given how intense the competition has become in this space. As the demand for GenAI-driven solutions grows, new providers and products of computation are—despite serious IP and capital barriers—entering the market, contributing to a more diversified and mixed hardware offering. One such example is AMD's recent release of its MI300X chips, designed for data center GenAI workloads. Hyperscalers looking to offer GenAI inference as a service do not want to rely on companies such as Nvidia and are increasingly designing their own specialized chips optimized for inference. To name a few, Google has developed its own TPUs, and Microsoft introduced the Maia 100 chip. As a result, dependency on general-purpose GPUs (mostly Nvidia's) is expected to decrease, and the total quantity of computing supply for GenAI model inference may well increase beyond what our model assumes.

Breaking points beyond computing hardware

The overall scenario we have described in some detail forces the following conclusion: Even if the computational intensity of GenAI inference continues on its exponential growth path, and even if adoption rates worldwide turn out to be faster than any comparable precedent would indicate, and even if the intensity of use grows as fast as data traffic has done during the smartphone revolution, nevertheless the "regime" of affordable, widely available computing power looks unlikely to break.

As we stated at the outset, it is impossible to confidently anticipate adoption patterns for novel, exponential technologies, many dimensions of which (like agentic workflows) lack suitable historical parallels. Still, we believe our model does give a sense of just how aggressive GenAI demand growth would need to be for the current regime to break if our bullish assumptions seem to be consistent with it.

So, what *else* might break the regime? There are at least three possibilities that are worth acknowledging.

One is an explosion in consumer demand for inference. The scenario that comes to mind might be one of mass adoption of multimodal models to generate the most compute-intensive content (video) for social media, for example. How could compute supply keep up if the more than 30 million videos uploaded daily on TikTok were generated using Sora? This scenario would indeed exert enormous pressure on computing supply because video generation is several orders of magnitude more compute-intensive than text generation. But we think consumer demand is unlikely to be a regime-breaker. What would be more likely

[12] Chip-on-Wafer-on-Substrate (CoWoS) is a 3D packaging technology allowing for the integration of multiple chips onto a single wafer. This approach improves performance, reduces size, and increases the overall efficiency of semiconductor devices.

[13] High Bandwidth Memory (HBM) is a computer memory interface and is primarily used in graphics cards and high-performance computing applications. It features more bandwidth, while consuming less power compared with traditional memory.

is that inference would become increasingly costly to consumers. One mustn't underestimate the price elasticity of demand for a product that most consumers continue to access for free.

A second possibility would be the case of a supply chain disruption impacting hardware production. The supply chain for computing hardware is geographically fragmented and has multiple chokepoints: While US-based Nvidia designs most high-end chips, it is Taiwan-based TSMC that manufactures them using equipment exclusively produced by Netherlands-based ASML, and so on. Geopolitical tensions, trade restrictions, and/or sanctions could severely affect the production and distribution of critical components.

Finally, and most importantly, the regime could break due to insufficient energy supply. Recall that our model focuses on data center–based GenAI inference, and data centers are very power-hungry indeed. To assess what our bullish demand model entails in terms of energy requirements, we calculated the terawatt-hours (TWh) per year needed to power all of the Nvidia A100 GPUs that are used in any given year for GenAI inference.[14] For the first year of our model, the power needs are high but reasonable, adding up to ~40 TWh per year, somewhere between 10% to 13% of today's global data center energy consumption (300 to 400 TWh). But once agentic workflows take off, energy quickly becomes a binding constraint absent substantial gains in energy efficiency.[15]

So, energy, not computing hardware, may very well be what breaks the "regime" of affordable compute in a scenario of bullish GenAI adoption. Nevertheless, there are signs that energy efficiency gains are possible, even likely. For instance, there already are advances in new hardware like Etched's Sohu, designed for specific tasks such as inferencing, offering superior performance with significantly lower power consumption than traditional GPUs. Beyond hardware, data centers themselves are becoming more energy efficient by optimizing cooling systems (e.g., direct-to-chip liquid cooling or immersion cooling) and leveraging AI itself to reduce energy requirements. Data centers are also exploring paths to on-site energy generation, which would alleviate pressure on energy transmission networks.

***

If our analysis is right and computing supply will meet even the most bullish GenAI inference demand scenario, computing is likely to remain affordable and widely available in the coming years. The challenge for most businesses is therefore not preparing for a world of computing power scarcity, but in fact the very opposite—getting ready to seize orders-of-magnitude more affordable compute to secure novel forms of competitive advantage.

[14] For the purpose of the calculation, we assumed Nvidia A100 Tensor Core GPUs with a performance efficiency of FP8 performance with about 500 TOPS/400 watts.

[15] While GenAI adoption may indeed lead to broader environmental sustainability challenges, those concerns are separate from the topic of this analysis—namely, whether economic or technological factors could disrupt the current model of affordable computing power, all else being equal.

About the Authors

Azeem Azhar is the founder of Exponential View; Executive Fellow at Harvard Business School and a technology investor.

François Candelon is a partner at the private equity firm Seven2 and the former global director of the BCG Henderson Institute.

Riccarda Joas is a consultant at BCG and an ambassador at the BCG Henderson Institute.

Nathan Warren is a senior researcher at Exponential View.

David Zuluaga Martínez is a senior director at the BCG Henderson Institute.

The authors would like to thank Leonid Zhukov and Meenal Pore, from the BCG Henderson Institute, for their contributions to this piece.

To learn more about our research, visit Exponential View *and the* BCG Henderson Institute.