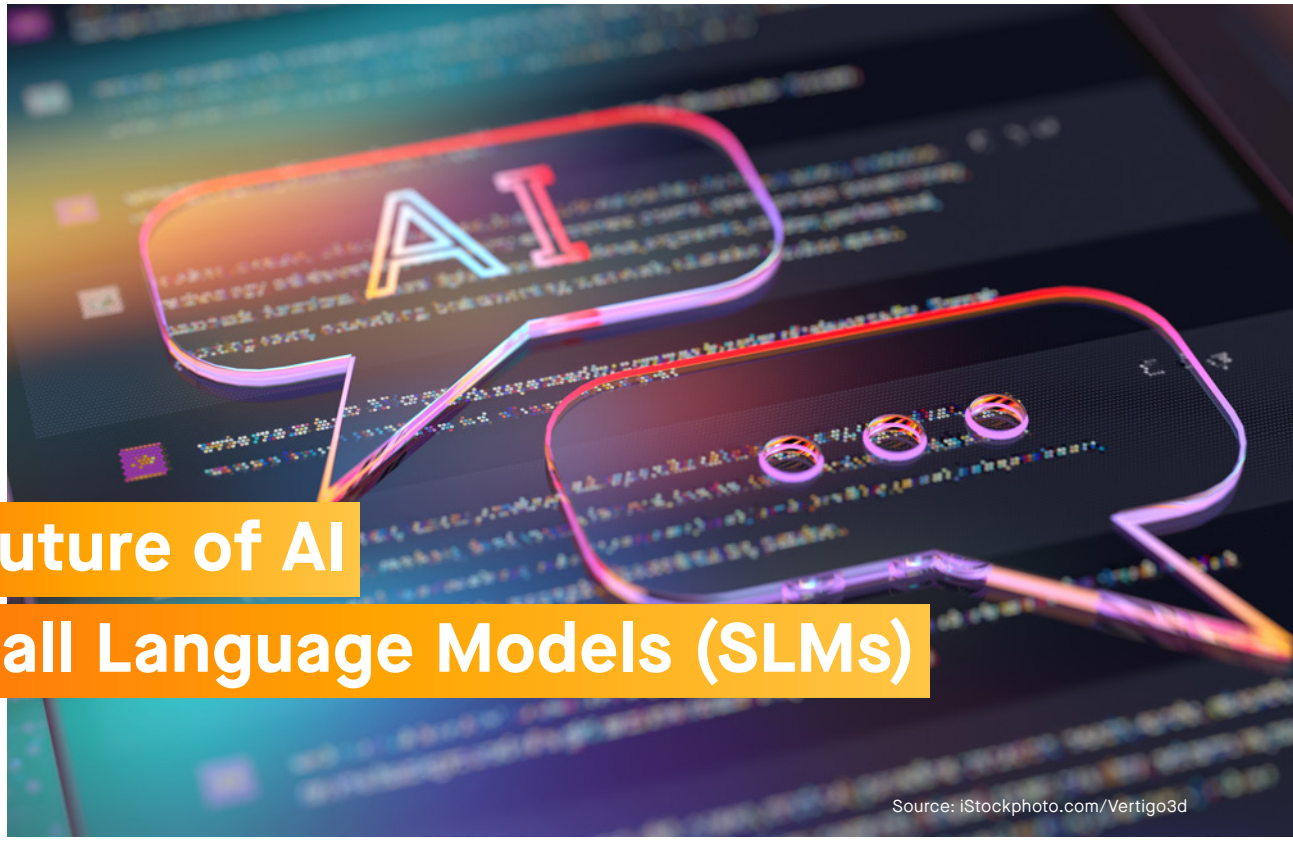


Source: iStockphoto.com/Vertigo3d

The future of AI is Small Language Models (SLMs)

It seems there has been a miscalculation within the AI industry. The industry assumed that progress required building ever larger language systems capable of tackling every conceivable task. These Large Language Models (LLMs) became the default solution, attracting immense investment and attention. In reality, most organisations do not need a Swiss Army knife when they are looking for a scalpel.

Charles-Henry Monchau, CFA, CAIA, CMT
Chief Investment Officer
charles-henry.monchau@syzgroup.com

Assia Driss
Syz Research Lab Team Coordinator
assia.driss@syzgroup.com

Syz PRIVATE BANKING

Introduction

A language model can be described as a system trained to learn the statistical structure of written text. During training, it analyses large volume of data and estimates the probability of each possible next word given the words that came before it. When the model receives an input sentence, it uses these learned probability distributions to select the next word that best fits the context.

A Large Language Model (LLM) applies this same mechanism, but it is built with a massive number of parameters, which allows it to capture a broader range of patterns and adapt to a wide range of subjects and tasks. This versatility, however, comes with a cost. Training and operating such models require significant computing power and specialised infrastructure.

A Small Language Model (SLM) follows the same principle but uses far fewer parameters. It is concentrated on more specific domains or functions, and it is thus lighter to run and easier to integrate into existing systems. Its compact design allows it to operate on modest hardware, to be fine-tuned quickly and to offer predictable performance for well-defined tasks.

The conventional wisdom has long held that AI progress means building ever-larger models that require massive computing resources and close ties to major cloud providers. SLMs are proving otherwise. In addition to being far cheaper and more efficient for specific tasks, these models give companies greater autonomy, stronger control over their data, and more flexibility, thereby reducing costly dependence on hyperscalers.

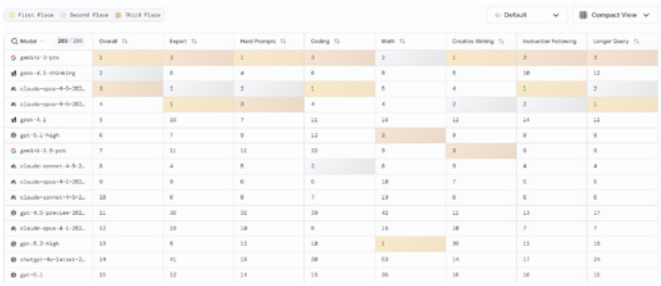
Large Language Models

LLMs are the result of decades of progress in natural language processing and machine learning, and they have been central to the acceleration of technological advances seen in recent years. They are now widely accessible through platforms such as OpenAI's ChatGPT, Google's Gemini, Microsoft Copilot, and Anthropic's Claude.

LLMs are basically deep learning systems trained on enormous amounts of text, giving them the ability to interpret and generate natural language across a wide range of tasks. They are built on a neural network architecture called a transformer, a design that excels at processing sequences of words and capturing relationships across long stretches of text. During training, an LLM is fed massive amounts of text, from books, articles, websites, code and more. The model "learns" by assigning probabilities to sequences of words, developing an internal statistical understanding of language. When given a prompt, the model uses this knowledge to predict the next word (or token) repeatedly until it generates a coherent response.

What distinguishes LLMs from simpler language models is size. Their internal parameters, the numerical values that encode all learned patterns, number in the billions or even trillions. Because of this scale and architectural complexity, they are especially useful when tasks require broad general knowledge, deep context awareness, or flexible handling of varied inputs. They can shift from drafting a marketing copy, to summarising research, to writing computer code, to engaging in a nuanced conversation. When equipped with agentic capabilities, they can even carry out multi-step tasks with a degree of autonomy that previously required human intervention.

But this power comes at a cost. LLMs require substantial computational resources. They are often hosted in cloud infrastructure rather than local devices and their operational costs rise quickly when they are used at scale. The versatility they offer is substantial, but so is the infrastructure required to sustain it.



Source: LMArena

SLM: When less is more

SLMs apply the same predictive principles as their larger counterparts, but with a fraction of the parameters, usually under 10 billion parameters. This reduction is not a limitation. By focusing on narrower domains and well-defined tasks, SLMs become lighter, faster and far easier to deploy.

Many SLMs can run on a laptop, an edge device or a local server without relying on the heavy cloud infrastructure that LLMs require. This local execution offers advantages such as reduced costs, predictable performance, and increased control over data, since the information never leaves the user's environment.

Their efficiency also makes them highly adaptable. While fine-tuning an LLM can take weeks and substantial Graphics Processing Unit (GPU) resources, an SLM can often be adjusted in hours or days on a single high-end GPU.

Despite their smaller size, modern SLMs deliver impressive capability. Models such as Google's Gemma 2 (2 billion parameters), Microsoft's Phi-3 (3.8 billion parameters), Meta's Llama 3.1 (8 billion parameters), NVIDIA's Nemotron Nano (9 billion parameters) or OpenAI's GPT-4o mini (number of parameters not disclosed), demonstrate that carefully optimised architectures can outperform much larger systems on specialised tasks, from code generation to reasoning benchmarks.

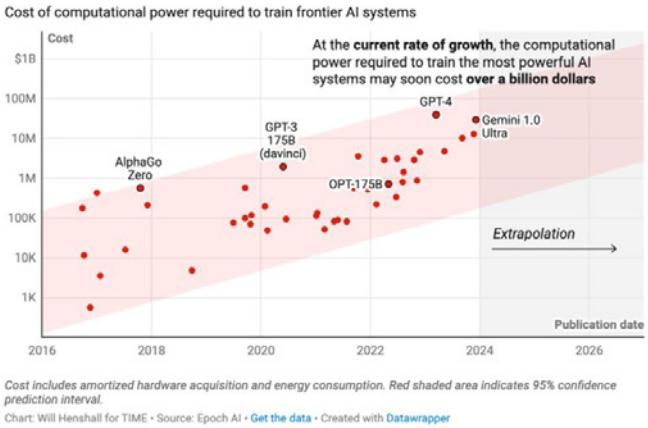
Recently, Microsoft introduced Fara-7B, an experimental small language model designed to run directly on a user's computer. It is described as the company's first agentic SLM built specifically for local operation, with the ability to control system inputs such as the mouse and keyboard.

It runs on seven billion parameters, far below the scale of earlier LLMs like GPT-3, which contained 175 billion parameters. Microsoft reports that Fara-7B "achieves state-of-the-art performance within its size class and is competitive with larger, more resource-intensive agentic systems that depend on prompting multiple large models."

In many real-world workflows, instruction following, tool use, or repetitive domain-specific tasks, a compact model is sometimes not only sufficient but often preferable, especially where computing resources are constrained. Systems such as autonomous vehicles or satellites operate under strict limits on processing power, energy consumption, and network access. In these settings, large models are simply impractical. Small language models, by contrast, can run directly onboard, enabling local decision-making without reliance on continuous cloud access.

Training a GPT-4-class model is estimated at over $100 million, with Gemini Ultra potentially reaching $191 million. Even adapting LLMs to specific domains can require tens of thousands of dollars in GPU time. By comparison, SLMs can often be trained and fine-tuned for a few thousand dollars. The difference is even more striking at inference. GPT-4 is priced at approximately $0.03 per 1,000 input tokens and $0.06 per 1,000 output tokens, resulting in an average cost of $0.09 per query. By contrast, an SLM such as Mistral-7B costs around $0.0001 per 1,000 input tokens and $0.0003 per 1,000 output tokens, or $0.0004 per query, a reduction by a factor of 225. At scale, across millions of requests, this cost gap materially affects operating costs and profitability, without even factoring in self-hosting expenses.

The cost of the computational power required to train the most powerful AI systems has doubled every nine months



Source: Will Henshall for TIME, Epoch AI

SLMs can then open access to use cases that would otherwise be out of reach. Schools, non-profit organisations and small businesses can deploy them for targeted tasks without facing prohibitive costs. In practice, models such as Microsoft's Phi-3 are already being used to support agricultural information platforms in India, delivering guidance to farmers even in regions with limited connectivity.

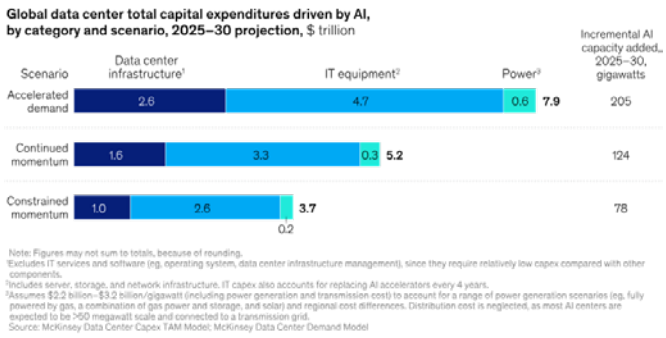
However, the efficiency of small language models comes with trade-offs. Their reduced size limits their ability to generalise across unfamiliar or loosely defined tasks, and they tend to struggle when problems require broad knowledge or deep, multi-step reasoning, affecting results on benchmarks. SLMs can also inherit biases from their training data, including biases originating from larger models. And like all generative systems, they can produce confident but incorrect outputs.

Consequences on Hyperscalers: The AI industry has backed the wrong horse

Hyperscalers have been pursuing a strategy built around scale, operating on the belief that ever-larger models and ever greater computing power would determine long-term advantage. The progression of flagship models reinforced that view. GPT-3, with 175 billion parameters, was widely viewed as a breakthrough in 2020, GPT-4, reportedly containing 1.8 trillion parameters, pushed expectations even further. The industry aligned itself with this trajectory and invested accordingly, rushing to build infrastructure before assessing the needs of real-world applications.

According to McKinsey estimates, total spending on AI infrastructure could reach between $3.7 and $7.9 trillion by 2030. In the second quarter of 2025, 98% of the $82 billion spent on AI infrastructure was directed toward servers, with 91.8% of that flowing into GPU- and XPU-accelerated systems. Hyperscalers and cloud builders accounted for 86.7% of total spending, or roughly $71 billion in a single quarter. Capital became heavily concentrated in highly specialised, energy-intensive hardware designed to train and operate massive models. Yet, the majority of enterprise applications simply do not require this level of capacity.

Capital investments to support AI-related data center capacity demand could range from about $3 trillion to $8 trillion by 2030.



Source: McKinsey & Company

According to a recent paper by NVIDIA Research, hardly a marginal voice in AI, "Small Language Models are the Future of Agentic AI," multi-agent systems show that between 40% and 70% of everyday tasks can be executed by SLMs without loss of effectiveness. In NVIDIA's words, "small language models are sufficiently powerful, inherently more suitable, and necessarily more economical for many invocations in agentic systems." Many agent-based applications today rely on models that are far larger than the tasks require. Replacing those heavyweight

systems with SLMs can reduce costs by up to 20 times while preserving performance across most workflows.

Despite their advantages, SLM adoption has been slower than expected. NVIDIA points to several structural reasons. Years of heavy investment have locked organisations into LLM-centric infrastructure, and industry benchmarks continue to reward scale, reinforcing the idea that bigger is better. The result is that although SLMs are often more practical and economical, the ecosystem remains shaped by large, cloud-based systems.

NVIDIA's paper goes beyond diagnosis and outlines a path forward. It advocates moving from monolithic LLM agents to modular, task-specific SLM capabilities, fine-tuned for real-world use and deployed locally where possible. The end state is a hybrid approach, where SLMs handle narrow, repetitive workloads and LLMs are reserved for tasks that genuinely require broad reasoning or open-ended interaction.

Conclusion

Hybrid architectures are becoming the norm. Small language models handle routine, well-scoped tasks efficiently. Larger models are reserved for complex queries that require broader reasoning or flexibility. The choice between small and large models is not about which is better, but about which is appropriate. Over time, effective systems will be defined less by scale and more by how precisely each model is deployed.

Welcome to Syzerland®

For further information

Banque Syz SA
Quai des Bergues 1
CH-1201 Geneva
T. +41 58 799 10 00
syzgroup.com

Charles-Henry Monchau, CFA, CAIA, CMT
Chief Investment Officer
charles-henry.monchau@syzgroup.com

Assia Driss
Syz Research Lab Team Coordinator
assia.driss@syzgroup.com