FORBES > LEADERSHIP > CMO NETWORK

# Small Language Models – More Effective And Efficient For Enterprise AI

**Daniel Newman** Contributor

*Exploring Cloud, AI, Big Data and all things Digital Transformation.*

Follow

0

Oct 26, 2024, 10:22pm EDT

Silhouetted office workers replaced by computer code. GETTY

Frontier models in the billions and trillions of parameters have been a focal point of the past two years as generative AI enthusiasm has continued to grow steadily finding its way into our apps, devices, and businesses–with new tools and use cases coming to market almost daily.

We also know that the rapid growth of large AI models for language, voice, and video is putting notable stress on resources, which has ignited a renaissance of interest in nuclear power as hyperscalers like Microsoft, Google, and AWS have all made sizable commitments to nuclear to support hundreds of billions of data center infrastructure build out expected over just the next few years.

And while models in the hundreds of billions and trillions of parameters like those developed by researchers at OpenAI, NVIDIA, Google, and Anthropic are at the cutting edge, we also know these power-hungry next generation models are often far more powerful than what is needed for most use cases–kind of like driving a Formula 1 race car in the middle of rush hour traffic.

This is where smaller models that can be powered with less energy and compute horsepower come into play.

## NVIDIA NIM and IBM Granite 3.0 Provide a Glimpse into the Future of Enterprise AI

More and more we are hearing about small language models with hundreds of millions or sub 10 billion parameters that are highly accurate and consume substantially less energy and cost less per token.

This past March at its GTC Conference, NVIDIA launched its NIM (NVIDIA inference Microservice) software technology, which packages optimized inference engines, industry standard APIs and support for AI models into containers for easy deployment. Inherently, NIM can handle models that are larger than small language, but the idea of optimized container services with industry specific models and APIs that could be used for visualization, game design, drug discovery or code creation represent an instance where the compute, data, models, and frameworks can be greatly simplified while also reducing the amount of computational horsepower to run AI workloads. I see the partnership that was recently announced between NVIDIA and Accenture as a great example of the combination of compute, industry specific microservices, and expertise to enable faster adoption of AI in the enterprise.

Last week, IBM announced its newest Granite 3.0 models, which are a family of small language models that showed strong performance against the likes of Llama and Mistral smaller language models (7-8 billion Parameter). All three companies have developed flexible open-source options that can be tuned and optimized for business use cases performing incredibly well in areas like math, language, and code. While Llama has been a staple of the open source model development, IBM's rapid improvement is noteworthy and with the companies open source offerings that can be used in clouds like AWS but also can be leveraged on IBM's own watsonx platform, I see these advancements as an example where an enterprise focused company like IBM with its software, models, and large consulting could pursue a strategy of "AI for Enterprise" effectively given the complexity to solve a continuum of use cases that will often require more than just models, but deep industry expertise.

Where this all heads are a mixture of models and flexible infrastructure that enterprises can focus on outcome-based AI projects that serve to enable the next wave of technological advancement like agentic AI, assistants and automation, and digital labor at scale.

## Research Will Persist but the Future Will Be Smaller Models for Enterprise

The idea that a one size fits all model with trillions of parameters is the holy grail of enterprise AI falls flat on a number of different fronts–Most notably the energy consumption and cost per token for well-defined use cases that really only need a few billion parameters (at most) to operate are simply better off being executed on specialized smaller models that are tuned for specific business use cases. Furthermore, governing and dealing with a mountain of growing data security, privacy, and sovereignty issues will be easier when the data lineage is better understood and access to data is limited to only what is required versus larger models that require massive scale to address a plethora of use cases.

Furthermore, there is no question that we want to continue to research and build the world's most sophisticated AI that will help support economic growth and aid in solving complex problems. But, for enterprises the smaller language and foundation models will prove to be a better option for many business use cases and will enable AI to be deployed at scale in a way that is more sustainable and better fit for purpose all the while meaningfully reducing the cost of AI. A combination that shouldn't and won't be ignored by businesses looking to capitalize on the potential of generative and agentic AI solutions.

*Follow me on Twitter or LinkedIn. Check out my website or some of my other work here.*

**Daniel Newman**

Follow

I am CEO of The Futurum Group. I spend my time researching, analyzing and providing the world's best and brightest companies with insights as to... **Read More**

Editorial Standards Forbes Accolades

One Community. Many Voices. Create a free account to share your thoughts. Read our community guidelines here.