AI Next

edgeverve
An Infosys company

Driving the Language Model Edge for the Enterprise

Guruprasad NV
AVP and Senior Principal Technology Architect, Infosys Topaz

Swaminathan Natarajan
VP & Global Product Head, EdgeVerve

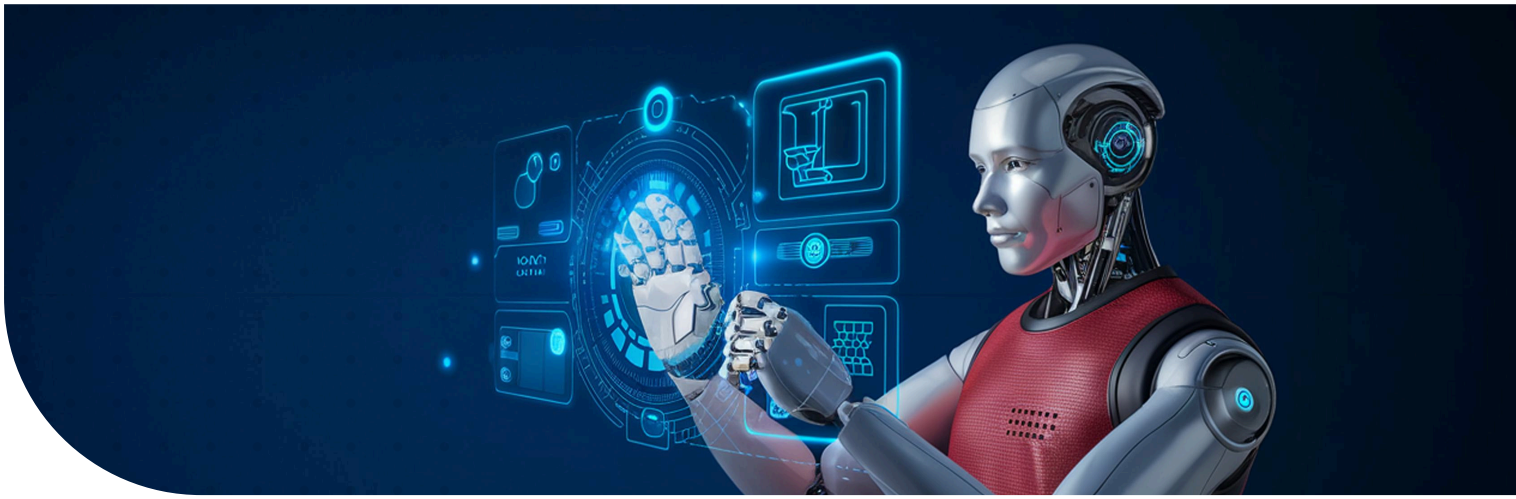
ABSTRACT

Every enterprise has felt it: the boardroom pressure to "**do something with AI**." In the rush, large language models became the tool of choice, promising transformation at scale. But behind the dazzling demos lurked harder questions -

- What happens when experiments need to become production systems?
- When costs balloon, compliance tightens, and trust is on the line, where does the real advantage lie?

Over the last two and a half years, we've seen enterprises everywhere jump headfirst into large language models. The pressure was real, and, in many cases, it came straight from the board. In these early days of exploration and innovation, organizations were liberal with experimentation, running proofs of concept in silos, working with multiple vendors, and trying ideas that looked exciting on paper.

That phase was important. It validated hypotheses and gave business leaders a taste of what was possible. And truth be told, LLMs are great in that initial phase when the goal is to experiment, to see if something works, to prove an idea. But the moment enterprises tried to take those ideas into production, the cracks began to show (see Fig. 1 on Page 3). While the outcomes were really exciting from an end user experience standpoint, financially, the ROI was not viable at all.



LLMs hallucinated, lacked the context of enterprise data, and raised compliance questions when sensitive information left organizational boundaries. Simple use cases demanded months of testing before they could be certified driving up both financial and time costs. **McKinsey** found that only one percent of companies could claim maturity on the deployment spectrum.

In short, enterprises discovered what many of us suspected early on: experimentation is easy, but scaling LLMs into production is hard.

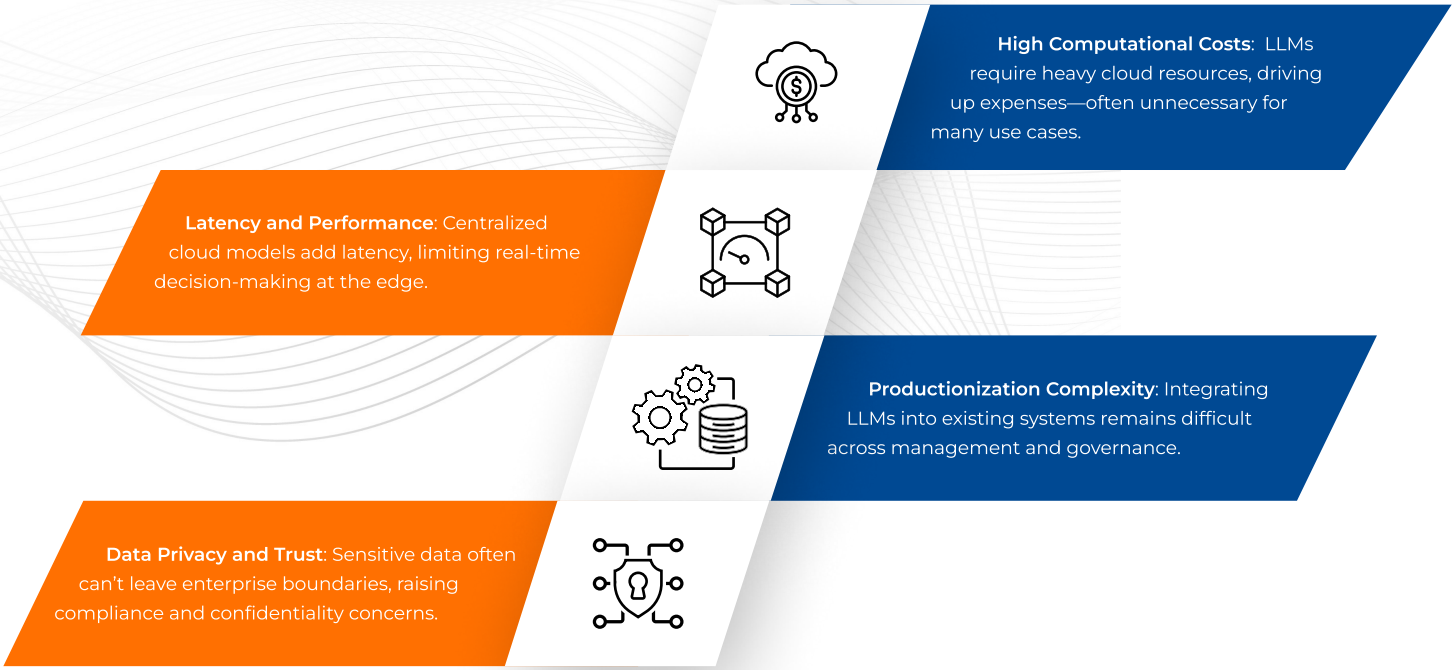


Fig 01: LLM challenges in the enterprise context

Why Purposeful Beats General at Enterprise Scale

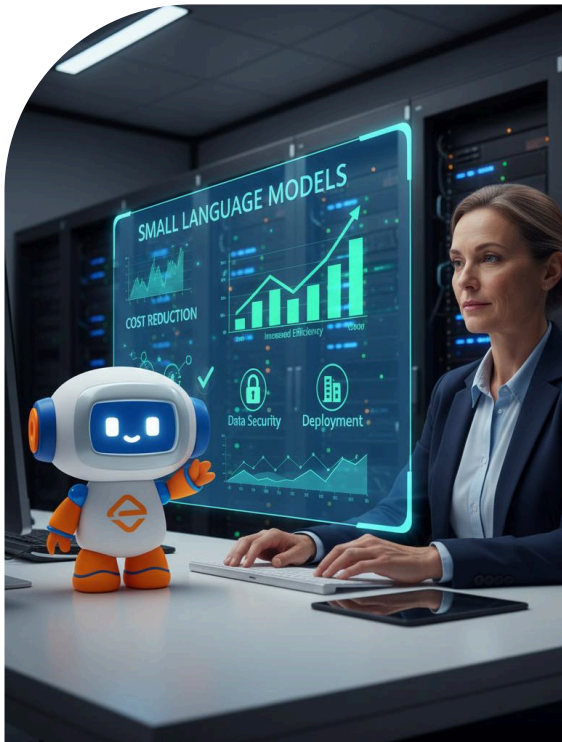
LLMs remain powerful, and as compute becomes cheaper their economics will continue to improve. But scaling them in their raw form inside an enterprise quickly became impractical. As generative AI evolved into token heavy agentic AI even small use cases started to consume a lot more tokens. There was a need to get leaner and more purposeful.

And so, questions emerged:

- Instead of throwing general-purpose LLMs at every problem, can we distill them?
- Can we build smaller, domain centric models tailored for specific business needs that can be hosted on much smaller infra, even locally, and still deliver accuracy and efficiency without runaway costs?

This line of thinking led to the rise of small language models, or SLMs. Unlike general-purpose LLMs that are trained on the open internet, SLMs can be trained or distilled specifically for domains such as financial services, healthcare, insurance, cybersecurity, or IT operations. They can be deployed within enterprise firewalls, ensuring data security. While they are not as powerful as LLMs, their size makes them cheaper to run, faster to fine-tune, and more predictable in behavior.

Most importantly, SLMs dramatically reduce the burden of testing and validation. As they are trained on enterprise data and patterns, they hallucinate less and stay within the bounds of domain logic. That makes them more trustworthy for business-critical applications.



We saw this firsthand when, in October last year, Infosys worked with NVIDIA to create our **first foundational small language model** to support some of our internal platforms such as Finacle. We noticed that in markets like India, Africa, and APAC, banks wanted the benefits of AI but had to contend with strict compliance environments and prohibitive cloud costs. Public LLMs were a non-starter. By training the SLM on our data and ecosystem inputs, tuning it for financial context, and making it light enough to run locally, we proved that enterprises could have AI capabilities inside their trusted platforms without breaching compliance or budgets. It was a strong demonstration of how a domain SLM could deliver where a general model could not.

Very quickly, the same pattern showed up in other engagements too. One of the world's largest active fund managers approached us with a clear need: automate the extraction of business rules from investment prospectuses so they could match opportunities to client profiles.

They began, like many others, with large public models. For feasibility, it worked beautifully - documents were parsed, patterns identified, rules surfaced. But when proprietary client data came into scope, the POC was stuck. Compliance leaders refused to allow sensitive information to flow through public endpoints. Finance teams ran the numbers and showed how token costs would balloon once scaled. To overcome these challenges, we moved the workload to a smaller, privately hosted Llama 3.2 model, with NVIDIA infrastructure and **EdgeVerve AI Next** orchestration layered on top. With this, the economics became viable, compliance concerns were addressed, and the use case moved into production.

This is not an isolated story. Whether in insurance underwriting, where third-party risk data must be combined with confidential client information, or in healthcare, where the volume and sensitivity of patient data makes token costs prohibitive, the pattern repeats: LLMs open the door, but SLMs make production possible.

From Feasibility to Viability with a Poly AI Approach

Across industries, we're seeing a familiar adoption lifecycle emerge. Most enterprises begin with LLMs because that's where the ecosystem momentum has been. The early proofs of concept demonstrate potential, but when teams start asking, "How do I scale this across my enterprise?" the questions change.

They want observability and control. They want predictable cost models. They want compliance certainty. And they want models that don't need endless rounds of testing just to be trusted with basic enterprise workloads. That's when SLMs enter the picture. Gartner predicts that in the next two years SLMs will be used at least three times more than LLMs.

However, LLMs are not going away. Business and technology leaders have realized they don't have to choose between ambition and pragmatism. The right approach is to use both: deploy LLMs where broad, general knowledge is required - for example, summarization across diverse document types or knowledge discovery across public data. Then deploy SLMs for high-value, contextual workloads where accuracy, privacy, and cost are non-negotiable such as customer service within regulated industries, domain-specific copilots, or platform-embedded AI functions.

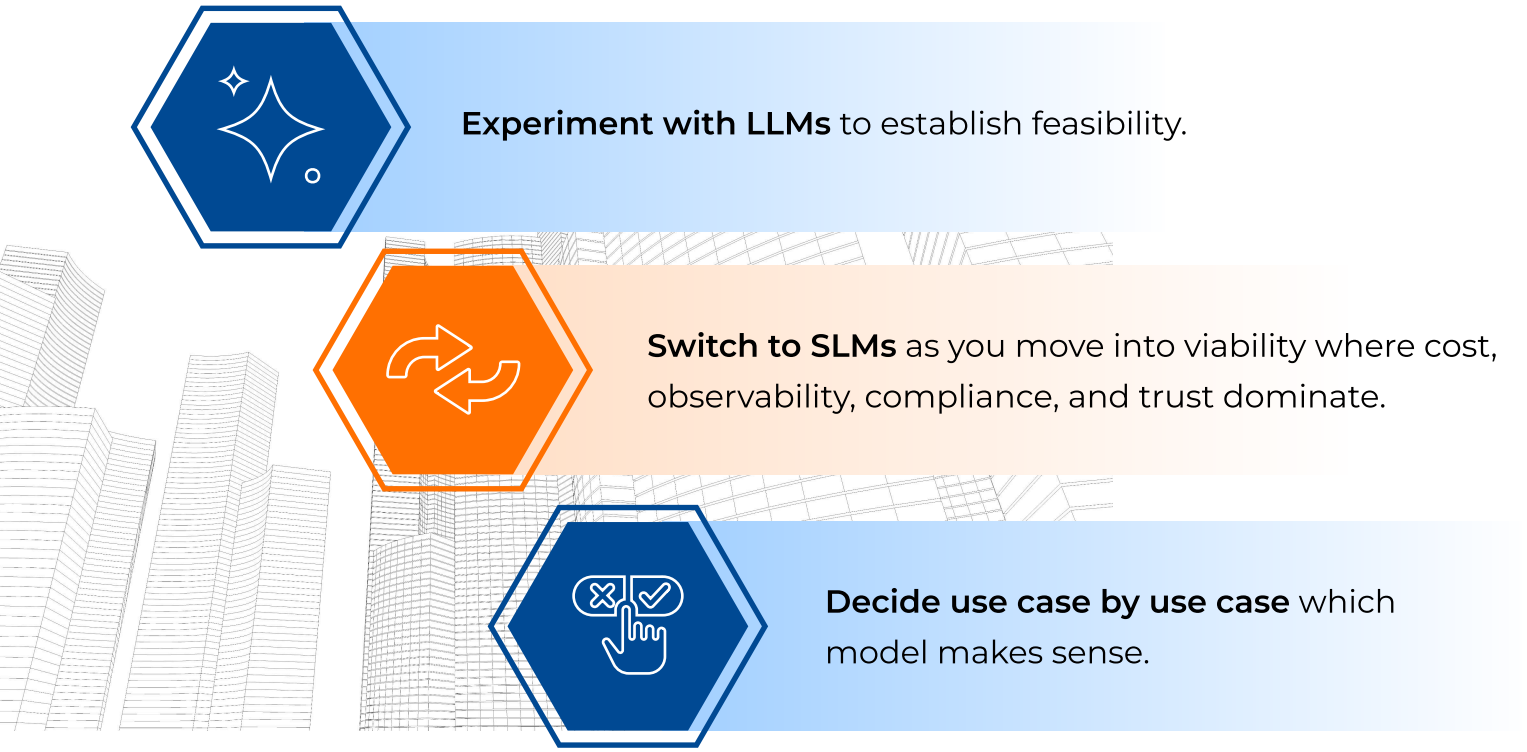


Fig 2: LLM challenges in the enterprise context

For instance, hospitals processing hundreds of thousands of diagnostic reports daily cannot afford the spiraling token costs of an LLM, nor can they allow patient data outside secure firewalls. For them, a smaller, domain-tuned model is the only viable answer. On the other hand, a pharma company analyzing diverse datasets occasionally may benefit more from the breadth of an LLM than from the narrow focus of an SLM.

Which is why, we prefer not worrying about early model choices. With the right capabilities (such as EdgeVerve AI Next Neural Connect), enterprises can begin their journey with any LLM, test feasibility, and then carry forward seamlessly into SLM territory once economics and compliance come into play. They can then encode the delta that an SLM needs to be enterprise ready.

The language model adoption story in the enterprise is no longer about "LLMs or SLMs." It's about choosing the right model for the right use case and building the architecture to manage both.

Why a Platform Approach Becomes Inevitable

But here's an added complication: the models themselves are evolving at breakneck speed. As experts say, "Data is permanent. Models are perishable." Gartner forecasts that global generative AI spend will reach $644 billion in 2025, with most of it going to infrastructure and model experimentation. Every few months, new variants of both LLMs and SLMs are released, with new capabilities and architectures that improve

performance and accuracy. At the same time, the workloads enterprises want to run on them are evolving from simple prompt-response interactions to agentic architectures where AI systems are reasoning, orchestrating, and executing across multiple steps.

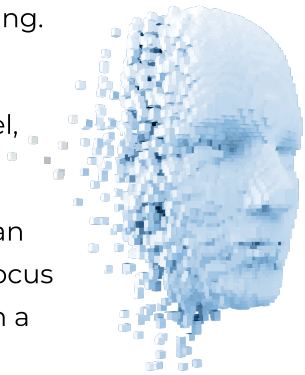
However, enterprises can't afford to rebuild their foundations every time a new model arrives. So, the enterprise problem is no longer just "which model do I choose?" The problem is: how do you build an AI architecture that can survive this pace of change? How do you adopt what's best today without locking yourself into something obsolete tomorrow?

This is why we believe a platform approach is non-negotiable. A platform abstracts the churn of models and allows organizations to plug in the right one without disruption. For a business user, that's what matters. They ask for a capability - summarization, extraction, speech-to-text - and they get it. Under the hood, if we swap a model for compliance reasons, or because a cheaper and better option came along, nothing breaks. That's how you scale AI in the enterprise without taking on unnecessary risk.

That's the principle behind Infosys Topaz SLMs and the Infosys AI Next platform working together. Topaz provides the domain-tuned SLMs that can be embedded directly into enterprise platforms. AI Next provides the unified layer to orchestrate these models with enterprise data, controls, and trust built in to insulate enterprises from volatility. By combining early access and ecosystem breadth with platform agility and enterprise focus, organizations can move beyond pilots to production - safely, scalably, and sustainably.

The Way to a Risk Managed Language Model Edge

One truth keeps coming back: models will keep changing. The pace is only accelerating. What works today may be obsolete tomorrow.



That's why our guidance to enterprises is simple: don't build around a single model, build around a lifecycle and a platform.

The real shift isn't just from LLMs to SLMs. It's toward platforms - architectures that can orchestrate between models, absorb the churn of rapid change, and let enterprises focus on outcomes rather than infrastructure. Build for that, and you give your organization a language model edge that is scalable, compliant, and economically sound.

To know how EdgeVerve AI Next Platform enables applied AI at scale, write to us at

contact@edgeverve.com

AI Next

EdgeVerve AI Next, part of Infosys Topaz, enables the scaling of Applied AI across the enterprise. Built from the ground up to leverage the power of Generative AI, this unified platform bridges silos in people, processes, data, and technology to drive transformation in business operations.
EdgeVerve. Possibilities Unlimited.
https://www.edgeverve.com/ai-next/

About EdgeVerve

EdgeVerve Systems Limited, a wholly-owned subsidiary of Infosys, is a global leader in developing digital platforms, empowering clients to unlock unlimited possibilities in their digital transformation journey. Our purpose is to inspire enterprises with the power of digital platforms, thereby enabling our clients to innovate on business models, drive game-changing efficiency, amplify human potential, and foster a connected ecosystem. Our comprehensive platform portfolio (EdgeVerve AI Next and TradeEdge) across Automation, Document AI, and Supply Chain helps inspire global enterprises to bridge silos in people, processes, data, & technology, discover & automate processes, digitize & structure unstructured data, and unlock the power of the network by integrating value chain partners. With a deep-rooted entrepreneurial culture, EdgeVerve's innovations are helping global corporations across sectors such as financial services, insurance, retail, consumer and packaged goods, life sciences, manufacturing, telecom, utilities, and more.
EdgeVerve. Possibilities Unlimited.
www.edgeverve.com