Futurum®

IBM

Leveraging Small Language Models for Enterprise AI: Benefits, Use Cases, and IBM's Approach

OCTOBER 2025

Introduction

The artificial intelligence landscape is undergoing a significant shift as enterprises move beyond the "bigger is better" mentality that has dominated recent AI development. While Large Language Models (LLMs) such as GPT-4 and Claude have captured headlines with their impressive general capabilities, a new category of AI models is emerging as more practical for many enterprise applications: Small Language Models (SLMs).

SLMs are specialized AI systems with fewer parameters compared to their large counterparts. Unlike LLMs that aim for high performance and broad general knowledge across countless domains, SLMs are designed for high-volume, low-complexity tasks or to be tailored for specific industries, tasks, or use cases This focused approach enables them to deliver superior performance in their target domains while requiring significantly fewer computational resources.

The AI industry's embrace of smaller, more specialized models reflects a growing recognition that enterprise success often depends more on targeted excellence than general capability. As organizations seek cost-effective AI solutions that don't compromise on performance or security, SLMs offer an alternative that delivers practical business value without the complexity and expense of frontier models. This shift represents a maturation of enterprise AI strategy, moving from experimental deployments to production-ready solutions that can scale across business operations.

Core Benefits of Small Language Models

Cost Efficiency and Resource Optimization

The most immediate advantage of SLMs lies in their dramatically reduced resource requirements compared to LLMs. Training an SLM typically requires fewer GPUs and shorter training times, making the initial development investment more accessible for enterprises. Where LLMs might require thousands of high-end GPUs running for months, SLMs can often be trained on more modest hardware configurations in weeks rather than months.

Beyond training costs, SLMs offer minimal inference costs that enable broader deployment across an organization. These models can run efficiently on single workstations or even mobile devices, eliminating the need for expensive cloud infrastructure for every AI interaction. This capability translates directly into reduced cloud computing expenses and lower ongoing operational costs as organizations scale their AI initiatives.

Perhaps most importantly, SLMs are significantly more practical to customize than large models. Organizations can fine-tune SLMs on their proprietary data at a fraction of the cost required for LLM customization, enabling domain-specific optimization that delivers superior business value.

Performance and Efficiency Advantages

SLMs offer powerful performance and efficiency advantages for enterprise AI, especially in practical scenarios where speed, cost and accuracy matter most. Unlike LLMs, which are often resource-intensive and designed for complex, broad-domain reasoning, SLMs excel in fast, targeted tasks. Their optimal architecture enables low-latency inference, which means users experience near-instant responses, improving customer satisfaction and enabling real-time applications such as chatbots, document summarization, and workflow automation.

For standard tasks like summarizing, categorizing, or initiating functions, SLMs deliver performance comparable to their larger counterparts, but without the unnecessary bulk. LLMs are often overkill for these purposes; their true advantage only becomes apparent in scenarios demanding expansive reasoning or navigating high levels of ambiguity, which are less common in typical business operations.

And where models have ben customized by the user for domain-specific tasks, SLMs tend to demonstrate superior performance in domain-specific tasks compared to general-purpose LLMs, which can be considerably more expensive to customize.

Deployment Flexibility and Accessibility

SLMs offer hybrid deployment options that can accommodate on-premises, cloud, or edge environments based on organizational requirements. While model size doesn't inherently dictate deployment flexibility, the reduced resource requirements of SLMs make diverse deployment scenarios more practical and cost-effective.

This flexibility proves particularly valuable for offline functionality in remote or secure environments where constant cloud connectivity isn't feasible or desirable. SLMs can operate independently, ensuring business continuity even when network connections are unreliable or prohibited by security requirements.

The compact nature of SLMs also enables integration with existing enterprise infrastructure without requiring wholesale technology overhauls. Organizations can incorporate AI capabilities into their current systems while maintaining scalability to meet varying demand patterns across different business units and use cases.

When and Where to Use Small Language Models

SLMs excel in several key enterprise scenarios where their focused capabilities and efficiency advantages provide clear business value.

Customer support and chatbots represent one of the most successful SLM applications. These models can provide 24/7 automated customer service while integrating seamlessly with company-specific knowledge bases. Unlike LLMs that are often cost-prohibitive to customize, SLMs can be trained cost-effectively to maintain consistent brand voice and messaging while accessing deep institutional knowledge about products, services, and policies.

Document processing and analysis benefits significantly from SLM's efficiency. These models excel at tasks such as document summarization and are just as capable of being when tailored to specific industry requirements, industry-specific content analysis that understands domain terminology and context.

Retrieval Augmented Generation (RAG) applications are particularly well-suited for SLMs, which can efficiently process and synthesize information from organizational knowledge bases to provide contextually relevant responses without the overhead of massive general knowledge models.

Edge AI applications leverage the compact nature of SLMs for mobile app and IoT device integration, enabling real-time processing without cloud dependency. This capability proves essential for applications requiring immediate responses or operating in environments with limited connectivity.

Agentic workflows benefit from SLMs' ability to automate tasks within specific domains while integrating with existing business processes. Indeed, a recent paper from NVIDIA Research argues that SLMs are a natural fit for agentic AI because most agent tasks are repetitive and narrow, not requiring general conversation abilities, agents typically use only a small subset of LLM capabilities through heavily constrained prompts. It argues that heterogeneous systems combining SLMs for routine tasks with selective LLM use for complex reasoning are optimal.

Decision Framework: SLM vs LLM Selection

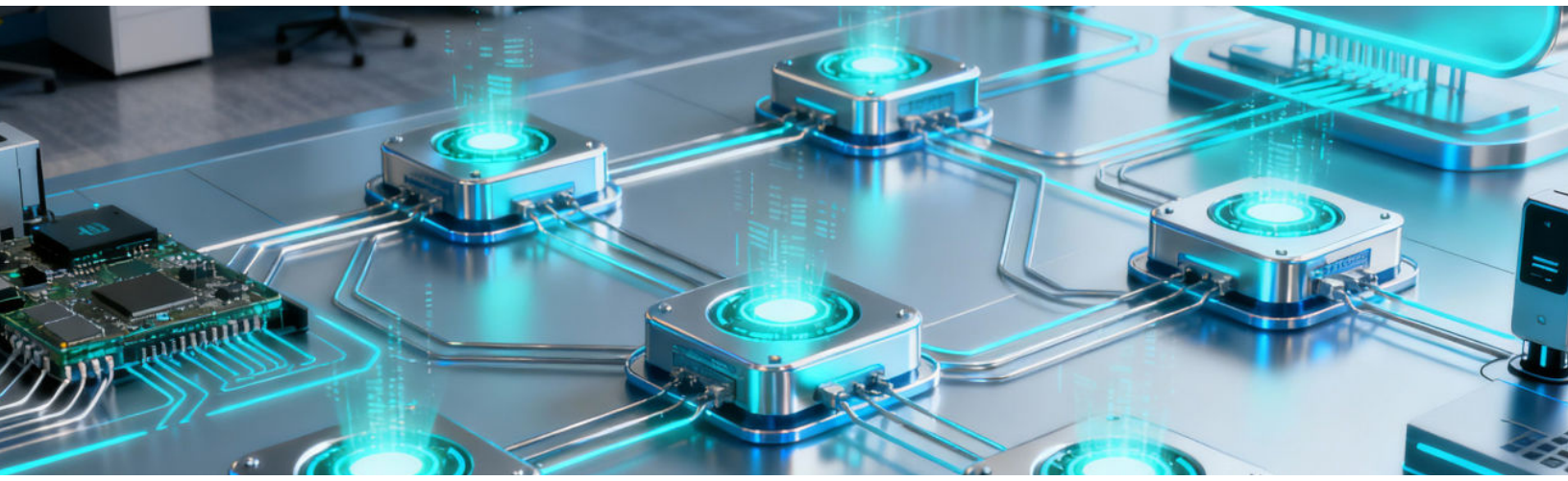
Organizations should prefer SLMs for well-defined, domain-specific tasks where the scope of required knowledge is clearly bounded (see Figure 1). SLMs enable faster deployment and iteration cycles, making them ideal for cost-sensitive deployments and situations requiring edge or mobile deployment. They excel in high-volume, low-complexity interactions where consistent, reliable performance matters more than broad reasoning capability.

Figure 1: Comparative Snapshot: SLMs vs LLMs

Dimension	Small Language Models (SLMs)	Large Language Models (LLMs)
Typical Parameter Count	100M–7B	30B–1T+
Training Compute	1–2 orders of magnitude lower	Thousands of A100-days
Fine-Tuning Time	Hours-to-days	Days-to-weeks
Inference Hardware	Mobile CPU/GPU, single workstation V100-32GB	Multi-GPU clusters
Domain Depth	High (specialized)	Broad (general)
Latency	Milliseconds	Seconds-to-minutes under load
OpEx	Low	High
IP/Compliance Risk	Lower (if self-hosting is an option)	Higher (API exposure)

Source: SLMs vs LLMs: What are small language models? – Red Hat, April 2025,

LLMs remain better choices when applications require broad general knowledge across multiple domains or complex reasoning that spans different fields of expertise. Organizations should accept longer implementation timelines for broader capability coverage when the use case demands creative content generation, research and exploration tasks, or complex problem-solving that benefits from extensive cross-domain knowledge.



IBM's Granite Approach: A Case Study in Enterprise SLMs

Granite Model Family Overview

IBM's Granite model family demonstrates the application of SLM principles in delivering enterprise-grade AI. These models, built on an open-source foundation with Apache 2.0 licensing, offer the transparency and control essential for enterprise customers. This licensing approach allows for customization and commercial use, avoiding the restrictions often associated with proprietary AI models.

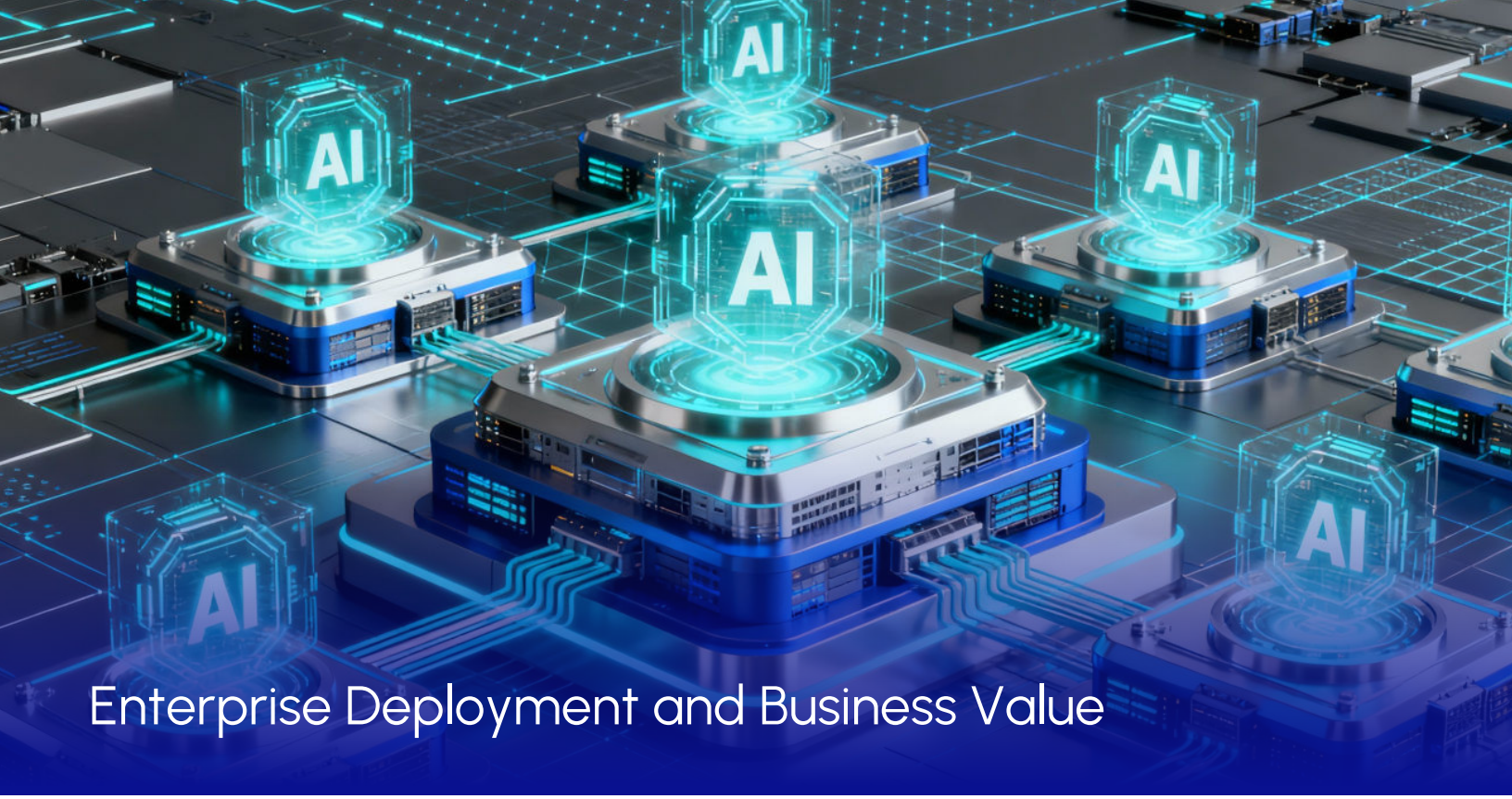
The Granite family encompasses a diverse range of model sizes designed for different use cases, ensuring organizations can select the appropriate model that balances capability with resource requirements. With multiple platform availability, Granite models can be deployed across various enterprise environments without vendor lock-in concerns.

Granite models demonstrate competitive performance metrics for enterprise tasks while incorporating multimodal and reasoning capabilities that extend beyond simple text processing. The models' transparency has earned Granite a top-five ranking on the Stanford University's latest Foundation Model Transparency Index, with clear documentation of training data and processes that enables organizations to understand and trust their AI deployments. Perhaps most significantly for enterprise adoption, IBM provides uncapped IP indemnification when using Granite models through watsonx.ai, offering legal protection that addresses one of the primary concerns enterprises have about AI deployment.

Granite 4.0 Innovations

The latest generation of Granite models introduces several innovations that address common enterprise AI challenges. Unconstrained context lengths enable extended document processing that can handle complex, lengthy documents without the traditional limitations that require document chunking or summarization before processing. And efficiency gains when running multiple long context sessions have resulted in around 80% reduction in GPU memory requirements when compared with similar models with different architectures.

Compact sizes optimized for flexibility and efficiency ensure that organizations can deploy Granite models across diverse infrastructure environments while maintaining enhanced architecture capable of handling complex workflows. Edge-optimized variants such as Granite 4.0 Tiny extend AI capabilities to resource-constrained environments without sacrificing essential functionality.



Enterprise Deployment and Business Value

Ease of Enterprise Adoption

IBM's Granite models streamline enterprise adoption through fine-tuning capabilities, enabling organizations to optimize performance using proprietary enterprise data for domain-specific applications. This customization integrates with existing tools and workflows, minimizing disruption during AI implementation. Additionally, multiple deployment options ensure compatibility with existing infrastructure investments.

Risk Management and Governance

IBM addresses enterprise risk concerns through Granite Guardian guardrail models, which provide automated risk mitigation for AI interactions. Enterprise-grade security controls and compliance-ready deployment options ensure organizations can meet regulatory requirements and internal governance standards. Furthermore, uncapped IP indemnification provides legal protection, enabling confident enterprise deployment without intellectual property liability concerns.

Business Impact and ROI

Organizations implementing Granite models can achieve domain-specific optimization at a fraction of traditional LLM costs, facilitating broader AI adoption across business units. This cost efficiency translates into improved operational efficiency and productivity gains that deliver measurable business value. Enhanced trust through transparency also enables faster executive buy-in, accelerating AI adoption timelines and reducing organizational resistance to AI implementation.

Conclusion: The Strategic Value of Small Language Models

SLMs represent a strategic evolution in enterprise AI deployment, offering cost efficiency, performance optimization, deployment flexibility, and enhanced governance that align with practical business requirements. Unlike the experimental deployments that characterized early enterprise AI adoption, SLMs enable production-ready solutions that can scale across business operations while delivering measurable return on investment.

The strategic implications extend beyond individual use cases to enable broader AI adoption across enterprise environments. By reducing barriers to entry and ongoing operational costs, SLMs democratize AI capabilities across business units that previously couldn't justify the expense and complexity of LLMs. This democratization accelerates digital transformation initiatives and enables organizations to realize AI's potential across a broader range of business processes.

Looking forward, SLMs will play an increasingly important role in the enterprise AI landscape as organizations mature from experimental AI projects to integrated AI-powered business operations. The combination of specialized performance, cost efficiency, and deployment flexibility positions SLMs as the foundation for sustainable enterprise AI strategies.

Organizations looking to adopt SLMs should consider a few key recommendations:

- Start with well-defined use cases: Focus on areas where domain expertise offers clear value.
- Evaluate total cost of ownership: This includes training, deployment, and ongoing operational expenses.
- Prioritize transparency and governance: Ensure these capabilities align with your organization's risk management requirements.

Begin by identifying high-volume, domain-specific processes that can benefit immediately from AI. As your organizational capabilities and confidence in SLMs grow, gradually expand their deployment. This measured approach fosters sustainable AI adoption, delivers lasting business value, and builds a foundation for more sophisticated AI implementations as the technology evolves.

Important Information About This Report

AUTHORS

Nick Patience
Vice President & Practice Lead,
AI Platforms | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

This document, including any supporting materials, is owned by The Futurum Group. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of The Futurum Group.

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT IBM

IBM is a global technology leader with a long history of advancing enterprise innovation across hardware, software, and services. Today, the company is focused on hybrid cloud, AI, and automation as the foundation for modern business transformation. A key part of IBM's AI portfolio is Granite, a family of open, state-of-the-art large language models designed for business applications. Granite reflects IBM's approach to trustworthy, domain-relevant AI that can be tailored for industry needs, enabling organizations to harness generative AI with greater transparency and governance. Through this combination of enterprise cloud, data, and AI capabilities, IBM continues to help clients modernize operations and drive measurable outcomes.

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337

© 2025 The Futurum Group. All rights reserved.