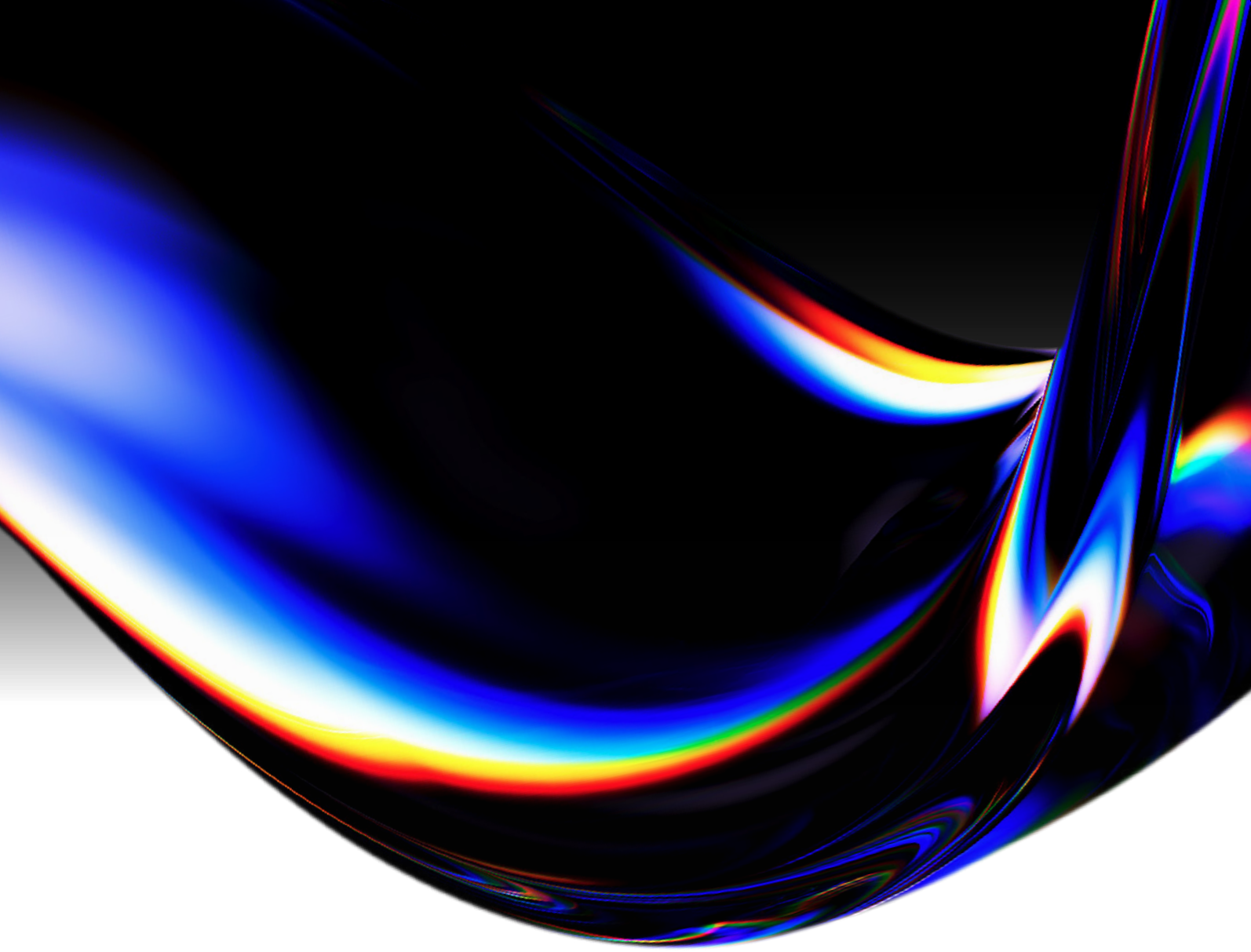

nagarro

Scaling AI in Telecom with Small Language Models

How small language models enable cost control, governance, and operational scale

nagarro

Table of Contents

Executive summary

Artificial Intelligence is rapidly reshaping the telecom industry by unlocking opportunities to automate operations, improve customer experience, and enable new digital revenue streams. However, as telecom operators move from isolated AI pilots to enterprise-wide deployment, they are encountering fundamental challenges related to cost, scalability, predictability, and governance. Token-based economics, latency constraints, data sovereignty requirements, and the risk of non-deterministic behavior limit the practicality of relying exclusively on Large Language Models at telecom scale.

This white paper argues that successful AI adoption in telecom requires a paradigm shift toward a hybrid, SLM-first architecture. Small Language Models, when fine-tuned on telecom-specific data and workflows, deliver domain-embedded intelligence with significantly lower cost, faster inference, and greater operational control. Large Language Models continue to play an important role, but primarily as a selective fallback for complex, ambiguous, or long-tail scenarios.

By combining fine-tuned SLMs, parameter -efficient techniques such as LoRA, and retrieval -based augmentation where appropriate, telecom operators can scale AI responsibly across OSS, BSS, network operations, and customer engagement. The paper presents a practical framework, supported by real-world industry examples, to help telecom leaders decide when to fine-tune, when to retrieve, and when to escalate to larger models; building AI systems that are fit for purpose, production-ready, and economically sustainable.

The Telecom AI challenge

The rapid advancement of AI, particularly Large Language Models, is reshaping industries by unlocking new sources of value, from enabling innovative revenue models to improving efficiency across the value chain. Telecom is no exception. AI is expected to significantly transform telecom operations by driving higher levels of automation across networks, platforms, and business processes.

Yet telecom operators operate under a unique set of conditions. Their massive scale, complex operational environments, and role as critical national infrastructure, demand AI solutions that are not only powerful, but also cost-efficient, predictable, and secure. While new LLM-driven use cases across network and business domains continue to emerge, deploying them at Telecom scale presents practical challenges. Token-based costs escalate quickly, and risks related to hallucinations, data security, and regulatory compliance become amplified.

These constraints call for a more balanced AI strategy. Rather than relying solely on large, general-purpose models, telecom operators increasingly require a hybrid approach. In this paradigm, Small Language Models play a central role by delivering domain-specific intelligence, lower operational costs, and greater control, while complementing larger models where broader reasoning is required.

Scaling AI in Telecom: Why Small Language Models

A real-world example of the impact of Small Language Models can be seen in how AT&T has scaled generative and agentic AI across its enterprise. Its internal platform supports over 100,000 users, processing over 750 million API calls and consuming billions of tokens daily.

As usage grew, the cost of relying solely on commercial Large Language Models became unsustainable. By adopting a hybrid AI approach, where fine-tuned Small Language Models serve as the core intelligence and large models are used selectively, AT&T leveraged proprietary telecom knowledge to drive efficiency at scale, achieving close to 90 percent reduction in operating costs.

Introducing Small Language Models (SLMs)

Small Language Models are lightweight versions of traditional language models designed to operate efficiently on resource constrained environments. While Large Language Models (LLMs) have up to trillions of parameters, SLMs typically fall in the range of a few hundred million to 10 billion parameters.

Small Language Models are task-focused AI models that are trained or fine-tuned to perform specific functions within a well-defined domain. Unlike large, general-purpose language models that aim to answer a wide range of questions, SLMs are optimized for accuracy, efficiency, and predictability within targeted use cases. Their smaller size enables lower inference costs, faster response times, and greater control over behavior, making them well suited for deployment in constrained environments such as on-premise data centers, private clouds, edge locations or even mobile devices. In telecom, where domain specificity, operational reliability, and data sovereignty are paramount, Small Language Models provide a practical foundation for scaling AI across networks, operations, and business workflows. The following table compares the impact of SLM vs LLM in the context of multiple dimensions.

Dimension	Small Language Models (SLMs)	Large Language Models (LLMs)
Model focus	Task- and domain-specific, optimized for predictable outcomes	General-purpose, optimized for broad reasoning and language understanding
Cost and scalability	Low, predictable inference cost, scales efficiently for high-volume telecom workloads	High, token-based costs that may increase rapidly at telecom scale
Accuracy and control	Higher determinism and lower hallucination risk within defined domains	Greater variability and higher hallucination risk in domain-specific atasks
Performance and latency	Fast response times suitable for real-time and operational use cases	Higher latency, especially for complex prompts and reasoning
Deployment and data control	Can be deployed on-premise, in private cloud, or at the edge with strong data control	Largely cloud-hosted with limited control over data residency
Reasoning capability	Optimized for narrow, well-defined reasoning patterns learned during fine-tuning; excels at rule-based, repetitive, and decision-oriented tasks	Stronger general reasoning and abstraction capabilities; better at handling ambiguous, multi-step, or loosely defined problems
Context length	Designed for shorter, focused inputs; performs best when provided with summaries ordecision-relevant context	Support significantly larger context windows, enabling processing of long conversations, documents, or cross-turn dependencies

Table 1: SLMs Vs LLMs

Popular SLMs

The Small Language Model ecosystem spans a wide range of model types, each optimized for different deployment patterns, performance needs, and operational constraints.

SLM Category	Optimized for	Representative SLMs (with params)	Where they shine (Examples)
General-Purpose Reasoning SLMs	Strong language understanding and reasoning with small footprints	Phi-2 (2.7B), Phi-3 Mini (3.8B), Gemma-2B (2B), Llama-3.2-3B (3B)	Intent detection, summarization, reasoning over bounded context, RAG answer synthesis
Agentic & Tool-Calling SLMs	Structured output, function calling, multi-step task execution	Qwen-2.5-3B (3B), Qwen-2.5-7B (7B), Mistral-7B (7B)	AI agents, workflow orchestration, OSS/BSS automation, API execution
Edge & On-Device SLMs	Low latency, low memory, offline or on-prem execution	TinyLlama (1.1B), SmolLM (135M–1.7B), OpenELM-3B (3B)	CPE/edge AI, on-device assistants, privacy-sensitive workloads
Multimodal (Vision / Speech) SLMs	Language combined with vision or audio understanding	MiniCPM-V (~2.8B), Qwen-VL-2B (2B), Whisper-Small (~244M)	Call transcription, document AI, image-based diagnostics, CV-enabled support
Domain-specialized SLMs	Optimized for a specific domain (code, speech, analytics)	DeepSeek-Coder-1.3B (1.3B), StarCoder-3B (3B), DistilBERT (66M)	Code copilots, NLP pipelines, classification, extraction

Table 2: Comparing the popular SLMs

Fine tuning SLMs

One of the key advantages of Small Language Models is the relative ease with which they can be fine-tuned. Fine-tuning can be applied either to adapt the model to a specific domain, such as telecom operations, or to optimize it for a well-defined task. Several approaches are commonly used to fine-tune SLMs, depending on the desired balance between performance and computational cost.

- Full fine-tuning – Full fine-tuning involves retraining all model parameters on new data and typically delivers the highest degree of customization, albeit at the cost of significant compute resources.
- LoRA (Low-Rank Adaptation) – Lightweight techniques such as Low-Rank Adaptation (LoRA) selectively fine-tune a small subset of model parameters, making the process faster and more cost-efficient.
- Adapters & prompt tuning - Adapter-based approaches and prompt tuning introduce additional layers or optimize prompt structures to guide model behavior, enabling targeted improvements with minimal changes to the base model.

LoRA fine tuning

Low-Rank Adaptation, commonly referred to as LoRA, is a parameter-efficient fine-tuning technique that has become especially popular for adapting Small Language Models to domain-specific and task-specific use cases. Instead of retraining all model parameters, LoRA introduces a small number of trainable low-rank matrices into selected layers of the model, while keeping the original weights frozen. This approach dramatically reduces the computational cost, memory footprint, and time required for fine-tuning, making it well suited for environments where resources are constrained or where models need to be adapted frequently.

In telecom contexts, LoRA enables rapid customization of SLMs for specific operational workflows without disrupting the stability of the base model. Operators can maintain a single foundational model and apply multiple LoRA adapters tailored to different domains such as network operations, customer care, or billing. This modularity allows teams to iterate quickly, deploy updates safely, and scale AI adoption across OSS and BSS functions while keeping inference costs low and governance manageable. As a result, LoRA has emerged as a practical enabler for production-grade AI in telecom, balancing specialization with operational efficiency.

RAG vs fine tuning

Retrieval-Augmented Generation (RAG) enhances a model's responses by dynamically retrieving relevant information from external knowledge sources such as documents, databases, or vector stores at inference time. The base model itself remains unchanged, and accuracy is improved by grounding responses in retrieved content. In telecom, RAG is particularly effective for the use cases that depend on frequently changing information, such as policy documents, product catalogs, pricing plans, network manuals, or standards. It allows operators to keep responses up to date without retraining models and provides traceability by linking answers back to source documents.

However, RAG primarily addresses knowledge freshness rather than behavioral specialization. While it can reduce hallucinations related to factual content, it does not fundamentally change how a model reasons, follows workflows, or executes domain-specific tasks. At large scale, RAG systems also introduce additional latency and operational complexity due to document ingestion pipelines, vector indexing, and retrieval orchestration.

Fine-tuning Small Language Models, on the other hand, directly modifies the model's behavior by training it on domain- or task-specific data. This approach embeds telecom knowledge, terminology, and workflows into the model itself. Fine-tuned SLMs are well suited for deterministic, repeatable tasks such as alarm classification, ticket summarization, intent detection, runbook guidance, and structured response generation. Because the intelligence is embedded in the model, inference is faster and more predictable, making fine-tuned SLMs ideal for high-volume and latency-sensitive telecom workloads.

The trade-off is that fine-tuning requires curated training data and periodic retraining as processes or domains evolve. It is best applied when tasks are stable and when consistent behavior and cost efficiency are critical.

Dimension	Fine-tuned Small Language Model	RAG-based LLM Model	General purpose Large Language Model
Task nature	Well-defined, repeatable, operational workfiows	Knowledge-centric tasks	Exploratory and open-ended reasoning
Scalability and cost efficiency	Low, predictable inference cost at scale	Moderate cost due to retrieval overhead	High and variable token-based cost
Latency and real-time suitability	Low latency, suitable for real-time operations	Added latency from retrieval layer	Higher latency for complex prompts
Domain accuracy and determinism	High precision, low variability	Grounded in retrieved content	Higher variability in domain-specific tasks
Hallucination risk	Lowest within trained scope	Reduced through grounding	Highest without strict controls
Knowledge freshness	Requires retraining for updates	Dynamically updated knowtedge	Dependent on model training cut-oft
Deployment and data control	On-prem, private cloud, edge ready	Primarity cloud-hosted	Primarily cloud-hosted

Table 3: Fine-tuned SLMs Vs RAG-based and General LLMs

Inspiration Story: Scaling GenAI in Telecom with Small Language Models

A leading telecom operator, AT&T, processes millions of customer service calls each year, generating vast volumes of recorded and transcribed conversations. To unlock actionable insights from this data, the organization initially adopted a Large Language Model–based approach to classify and analyze customer interactions across more than 80 service-related categories. While this solution delivered strong accuracy and measurable business impact, it quickly revealed critical limitations when deployed at scale, including high token-based costs, long processing times, and concerns around data privacy and security.

To overcome these challenges, AT&T adopted a hybrid model strategy centered on Small Language Models. Instead of relying on a single large model, the company distilled its AI workflow into a layered architecture that combined fine-tuned classifiers, domain-specialized SLM, and selectively used open-source LLM for the most complex cases. Most of the customer interaction categories were handled by lightweight, fine-tuned models, while only a small subset required larger-model reasoning. This approach reduced processing time from hours to minutes, increased transcript

Decision guide

As telecom operators move from AI experimentation to production, choosing the right model architecture becomes a strategic decision. While fine-tuned Small Language Models are particularly effective for operational and domain-specific workloads, they are not universally applicable. The following decision flow diagram helps clarify when an SLM-first approach can deliver the greatest impact.

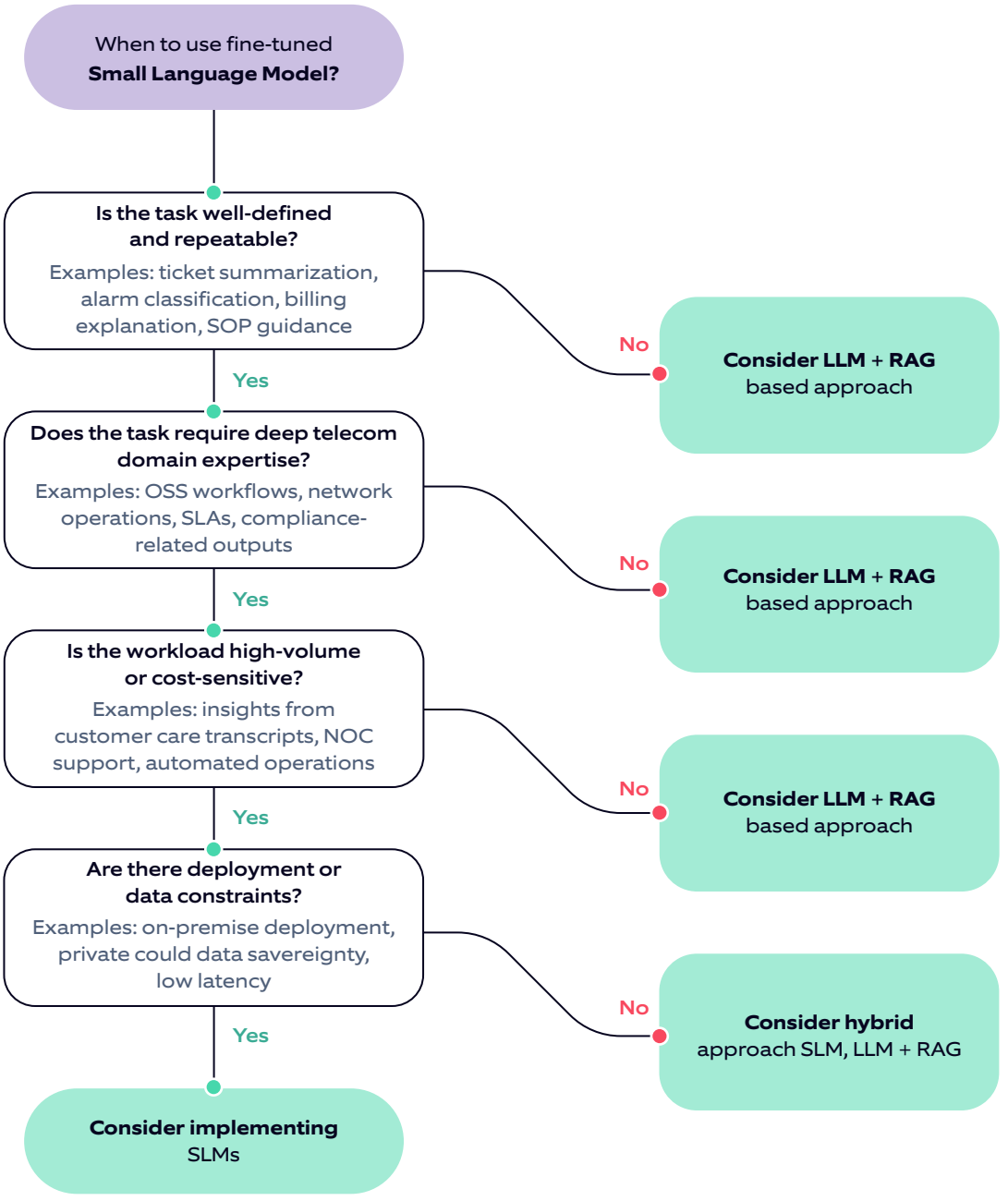

Figure 1: Decision flow for selection

Conclusion: Building Scalable and Responsible AI for Telecom

As telecom operators accelerate their adoption of AI, it is becoming clear that scale, control, and operational reliability matter as much as raw model capability. While Large Language Models have demonstrated impressive potential, deploying them indiscriminately across telecom environments introduces challenges related to cost, latency, governance, and data security. Telecom's unique role as critical national infrastructure demands an AI strategy that is deliberate, resilient, and designed for production from day one.

Small Language Models provide a practical and powerful foundation for meeting these requirements. By focusing on domain-specific intelligence, predictable behavior, and efficient deployment, SLMs enable telecom operators to move beyond isolated pilots and embed AI deeply across OSS, BSS, network operations, and customer engagement. When combined with parameter-efficient fine-tuning techniques and deployed within hybrid architectures, SLMs deliver much of the value of larger models while significantly reducing operational risk and cost.

The most successful telecom AI strategies will not be defined by a single model choice, but by thoughtful orchestration. An SLM-first approach, complemented selectively by larger models for complex reasoning, allows operators to scale AI responsibly, maintain governance, and continuously innovate. For telecom leaders, the path forward is clear: build AI systems that are fit for purpose, designed for scale, and aligned with the realities of telecom operations. Small Language Models are not just an optimization; they are the cornerstone of AI-native telecom transformation.

References

- https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/how-generative-ai-could-revitalize-profitability-for-telcos
- https://inform.tmforum.org/features-and-opinion/how-att-and-verizon-are-scaling-ai
- https://h2o.ai/content/dam/h2o/en/marketing/documents/2025/ATT-GenAI-CaseStudy-WhitePaper.pdf

Author

Vaibhav Nigam
Associate Director - Telecom
Nagarro

About Nagarro

Nagarro, a global digital engineering leader, helps clients become fluidic, innovative, digital-first companies and thus win in their markets. The company is distinguished by its entrepreneurial, agile, and global character, its CARING mindset, and its Fluidic Intelligence vision. Nagarro employs around 17,700 people in 39 countries. For more information, visit www.nagarro.com.