amdocs | make it amazing

AI Verticalization for Telco

Whitepaper

Table of contents

Abstract

This document explores the transformative potential of AI verticalization within the telecommunications industry. By customizing AI capabilities to meet industry-specific needs, companies can enhance efficiency, effectiveness, and trust. The document delves into the specific functionalities required for AI agents to support telecommunications companies, focusing on both telco-grade and enterprise-grade agents and their operations.

Key topics include the unique attributes and requirements of telco AI agents, such as understanding telco terminology, ontology, and context. Also covered is the integration of advanced technologies from Amdocs, NVIDIA, and AWS to deliver personalized customer experiences and robust operational capabilities. Additionally, practical frameworks for AI agents' personality engineering and customer experience personalization are discussed.

By providing a unified definition of telco-grade agents, this document provides readers with a comprehensive understanding of the critical functions and benefits of AI verticalization for telecommunications. It concludes with a practical example, illustrating the transition from theoretical frameworks to real-world implementation through a collaborative project undertaken by Amdocs, NVIDIA, and AWS.

The value of AI Verticalization

AI verticalization enhances agent design by tailoring capabilities to address industry-specific needs and functions. This approach ensures that agents acquire the domain-specific knowledge required to deliver more precise and relevant services. By focusing on a specific industry, the agent develops familiarity with its unique needs, empowering it to provide tailored solutions that improve efficiency and effectiveness.

Verticalized AI agents can quickly adapt to changes while staying accurate and reliable. This not only boosts customer satisfaction but also drives innovation in the industry. It allows for the creation of advanced, customized applications.

For example, verticalized agents can handle complex operational tasks, from customer service to network management, leveraging advanced machine learning algorithms and deep learning techniques to interact seamlessly with users. Furthermore, their analytics capabilities allow them to process and interpret vast amounts of data in real-time, providing actionable business insights.

Another key strength is their proficiency in managing end-to-end processes. By integrating with industry-specific systems and platforms, verticalized agents ensure smooth operations across the ecosystem. Unlike generic AI systems, they can autonomously initiate actions, monitor performance, and respond to dynamic changes. Additionally, their robust security protocols safeguard sensitive information, ensuring compliance with industry regulations.

Verticalized AI Agents adhere to the following principles:

- **Understanding industry terminology**
- **Developing specific ontology**
- **Enabling reasoning**
- **Maintaining the domain context**

By implementing these principles, verticalized AI agents can significantly enhance operational efficiency, customer satisfaction, and overall industry innovation.

Telco Grade Agents

The dynamic communications industry demands a unified framework for standardized telco-grade agents, especially given the market's complexity with numerous providers and varying standards. Effective standardization ensures consistent performance, interoperability, and efficiency, enabling agents to deliver high-quality services while following the industry's best practices.

What is a "Telco Agent"?

Telco GenAI agents are advanced AI systems that possess extensive knowledge and expertise in the telco domain. Training on telco data enables them to perform effectively across the entire industry ecosystem.

Using advanced AI technologies like machine learning and natural language processing, telco AI agents understand and interact with users naturally. They excel at handling large amounts of data, making smart decisions, and improving their performance over time. Their main strength is being able to act autonomously and being proactively, connecting to different systems, and communicating directly with customers and other agents to manage entire processes.

Verticalization of Telco Agents

Implementing telco AI agents begins by establishing a solid foundation rooted in industry expertise. Initially, intelligence is extracted to replicate the best behavior and provide the optimal customer experience tailored to telco needs, rather than merely offering 24/7 availability of bots. During this extraction process, different roles, their functions, the processes they follow, the systems they connect to, and the variations of these processes are identified alongside their performance metrics specific to telco operations.

Next, intelligence is applied by establishing the appropriate composable architecture and platform to support it. This foundational approach ensures that every aspect of telco operations is considered, enabling agents to deliver a hyper-personalized and visually impactful experience without the need for each customer to navigate complex systems individually. The application can serve various channels, from mobile apps to portals and voice calls, . Additionally, companies can introduce a multi-modal avatar with visuals and interface to complement telco-specific GenAI applications.

On the foundation level, an efficient ingestion pipeline is required, to ensure real-time processing of data (e.g. network domain). Large telco models, trained specifically on telco data, form the backbone of these agents, furnishing them with the necessary domain-specific knowledge. Furthermore, the agents must be adept at operating autonomously, facilitated through simulations. For instance, in telco networks, these autonomous agents can manage large volume of network data and optimization tasks, such as dynamically adjusting bandwidth based on real-time data, thereby enhancing network performance and reliability without human intervention. Notably, NVIDIA's recently announced an NVIDIA AI Blueprint for telco network configuration, includes real-time ingestion pipeline and validation tools.

a. Telco ontology

Ontology is a formal representation of knowledge within a specific domain, enabling consistent data interpretation. In the telecommunications industry, a well-defined ontology is crucial for ensuring that AI agents accurately understand and respond to various queries. Ontologies can be specific to a domain or even a particular task, providing a structured framework that enhances the functionality and accuracy of AI applications.

A telco ontology allows domain expertise to be introduced effectively. By leveraging in-context learning and large telco models (LTMs) - which are LLMs or SLMs trained specifically on telco data - reasoning is enhanced and flexibility is improved. These LTMs serve as the foundation for agentic workflows and reason based on the encoded knowledge, ensuring that all agents interpret data uniformly and deliver accurate and reliable information to customers. This structured framework allows for interoperable knowledge representation, enabling seamless communication and collaboration among different systems and agents, all while adapting to various contexts in order to provide more relevant responses.

b. Composable architecture & telco skills

Composable, modular and cloud-native architecture allows modularity and blending of deterministic and agentic capabilities, enabling reusability and control of the AI span of reasoning.

For real-time customer agents, hierarchical agentic layers are implemented. This involves a front manager router that triages incoming requests and multiple expert agents that reason and share the same state. This composable architecture allows for higher accuracy and reduced costs, as the AI is being asked to reason over a smaller scope. A skill is the most 'atomic' building block that plays a specific role. Within the composable structure there are agents of domains and their associated sub-agents, which formulate the skills, either simple LLM calls or more composite agents with tools. These skills are designed and implemented with specific instructions and prompts, aligning with the domain ontology. For example, skills like bill review, explain a charge, open dispute ticket, and usage details are classified and implemented as Care sub-agents oriented to the Care domain.

By implementing a composable architecture with hierarchical agentic layers, can optimize their real-time agents for higher accuracy and cost efficiency. This approach allows for superior service delivery while maintaining flexibility and scalability to evolving demands for orchestration and agent to agent communications, such as NVDIA AI-Q which enables communication and interaction across different agents. Cloud-native deployment enables telco-grade agents to deliver enterprise-grade production capabilities while rapidly transitioning from proof-of-concept to production. AWS's comprehensive data services organize and make proprietary telco data easily accessible, ensuring agents meet four critical requirements: adaptability to new use cases, scalability with high performance and resilience, seamless integration with proprietary data, and adherence to the highest standards in privacy, security, and responsible AI.

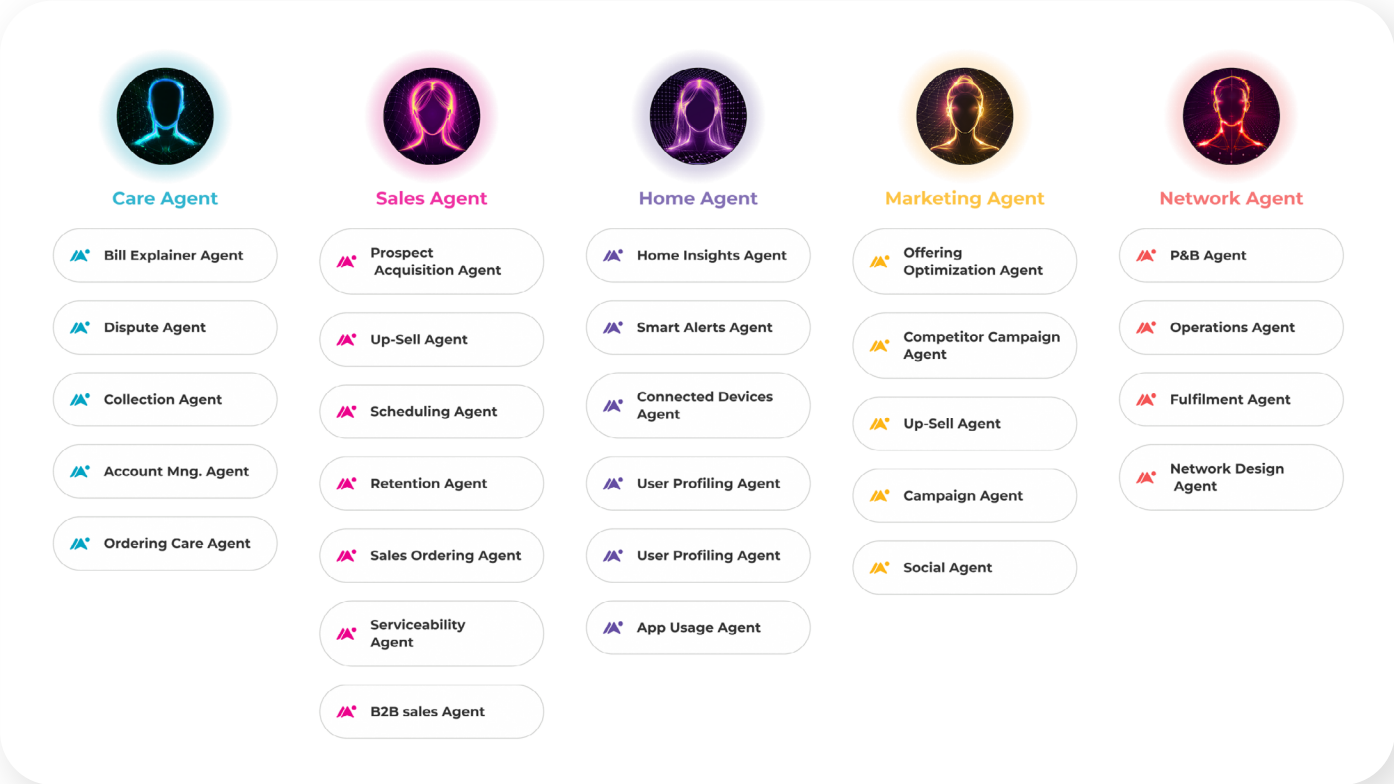

Figure 1 Each agent has a sub-agent which we call skill; each sub agent may be a simple LLM call or a more composite agent with tools

c. Telco reasoning

In the telecommunications industry, reasoning is a cornerstone capability for GenAI agents, enabling them to go beyond basic question-answering and handle complex, multi-step interactions across care, sales, network, and billing domains. Effective reasoning empowers AI agents to understand context across multi-turn dialogues, handle ambiguity in customer inquiries, navigate decision trees (e.g., diagnose network issues, plan recommendations) and synthesize information from multiple systems (CRM, billing, service status).

This is critical in telco environments, where workflows are often fragmented, customer intents span multiple domains, and real-time resolution is a business imperative.

For example, NVIDIA's Llama Nemotron model family introduces enterprise-optimized reasoning frameworks that can perform chain-of-thought analysis, weigh trade-offs, and adapt based on customer history and intent. These capabilities are especially valuable for orchestrator agents that coordinate across multiple specialized agents or back-end systems.

d. The critical role of digital twin

A Digital Twin serves as a sophisticated virtual model that mirrors real-world entities or systems, enabling telco-grade AI agents to operate with enhanced precision and reliability. This virtual representation is especially valuable in telecommunications, where it acts as a replica of complex networks, allowing agents to simulate scenarios, analyze data, and derive actionable insights without disrupting actual systems. By leveraging a Digital Twin, tTelco-grade AI agents can confidently execute decisions and manage intricate workflows, ensuring seamless service delivery and optimized performance. Foundational components such as data, analytics, and AI/ML drive the effectiveness of these Digital Twin implementations, making them indispensable for modern Telco operations.

e. E2E agentic experience

Since telco Agents are engaged in sensitive customer interactions, telcos should exercise extreme caution about how AI agents behave, respond, and adapt in real-time - across a wide range of customer contexts. A billing inquiry, Wi-Fi issue, delay in discounts or a technical service complaint may each require a different action, process invocation, tone of interaction and depth of explanation.

By providing a hybrid response - incorporating multi-modality in both visual elements, voice and text agents can create more natural and engaging conversations. This approach not only improves customer satisfaction but also reduces operational costs. Supporting content, such as images, diagrams, and infographics, can help clarify complex information quickly and effectively, leading to faster resolution of customer queries and more seamless interaction.

Enterprise-grade agents must be capable of making decisions and taking actions in real time, without waiting for human input. This shift from reactive to proactive service delivery is a key differentiator in modern customer experience. Rather than waiting for a customer to report a problem, agents can detect anomalies - such as a billing spike, service disruption, or usage pattern change - and initiate the appropriate workflow.

For example, in some telco environments, AI agents already identify potential billing issues before the invoice is even sent. If a customer's usage pattern deviates significantly from their norm, the agent can flag the anomaly, simulate the upcoming bill, and proactively reach out to the customer with an explanation or offer to adjust the plan. This not only prevents dissatisfaction but also builds trust and reduces inbound call volume.

Autonomous agents can also trigger upsell or retention workflows when they detect signals like contract expiration, churn risk, or eligibility for a better plan. These actions are not scripted - they're context-aware, data-driven, and executed in real time.

Due to the complex nature of Telco workflows, different types of agents can work together, each specializing in a specific area like billing, technical support, or sales. For example, collaborative agents divide complex tasks and build on each other's progress, making them well-suited for back-office operations. In more dynamic scenarios, a supervisor agent can coordinate responses across multiple domains, ensuring consistency and accuracy. For customer-facing services, a hierarchical team structure allows simpler requests to be handled quickly by lower-level agents, while more complex issues are escalated to advanced agents. This coordinated approach ensures faster, more accurate responses, and helps scale AI support to match the complexity of real-world telecom interactions.

However, customer interactions - especially those involving complex, emotional, or high-stakes issues - **still require human expertise**. AI agents must be able to recognize these situations and initiate a seamless handover to a human agent or supervisor. This ensures that customers receive the right level of support without friction or repetition.

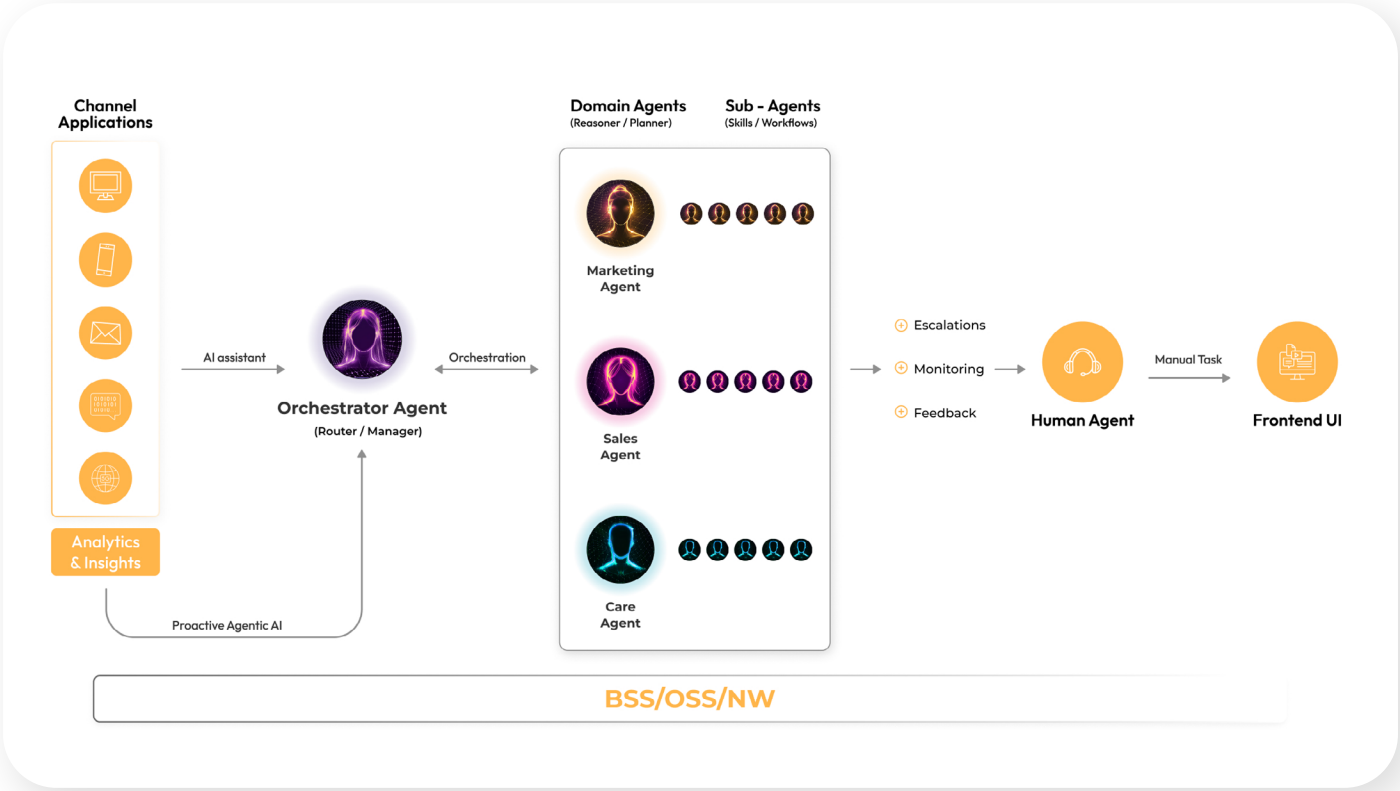

Figure 2 This illustrates the agentic experience using the composable architecture to achieve interactions between AI agents and human agents.

Enterprise-Grade Agents

Beyond telco-specific capabilities, enterprise-grade agents require robust horizontal attributes that span across functions – including brand engineering, personalization, trustworthiness, and autonomous operations. By implementing these features, enterprises can ensure that AI agents perform their internal and customer-facing functions in a way that strengthens customer relationships and confidence.

Brand engineering

As AI agents become the primary interface between companies and their customers, they are reshaping how brands are experienced across industries. Whether in banking, aviation, transportation, or telecom, these agents are no longer just service tools - they are the brand's voice, tone, and personality in action. Global research (Amdocs, 2025 "Brand new frontier: the agentic era research"), spanning 14 countries and over 7,000 consumers, reveals a striking insight: **80% of consumers trust AI agents to resolve service-related issues**, and **60% believe that AI agents can positively influence their perception of a brand** - when implemented thoughtfully. This trust, however, is not automatic. It must be earned through intentional design and consistent brand alignment. In brand engineering, voice and visual play a critical role:

Voice is more than just sound - it represents the first impression of your brand. Whether engaging with a bank, airline, telco, or transit provider, the AI agent's voice sets the tone for the experience. Customers care deeply about how AI agents sound, with preferences varying by age, gender, and region. For example, 67% of women prefer female-voiced agents, and many consumers favor voices matching their age or cultural background. AWS's Project NOVA and NVIDIA Riva play key roles in enabling real-time, personalized voice interactions at scale, providing lifelike, branded voices that respond with high performance and low latency.

Visual design is crucial for the perception and trust of AI agents. While some users prefer abstract or logo-based representations, others welcome expressive digital personas, avoiding the eerie "uncanny valley." With 49% of consumers wanting to customize agent characteristics like age, gender, and style, dynamic, context-aware visual agents become essential. BlueprintNVIDIA's Digital Humans for Customer is a suite of digital human technologies that bring digital characters to life with GenAI, enabling real-time animation, speech recognition and synthesis, enterprise information retrieval and conversation management that aligns with brand identity and tone.

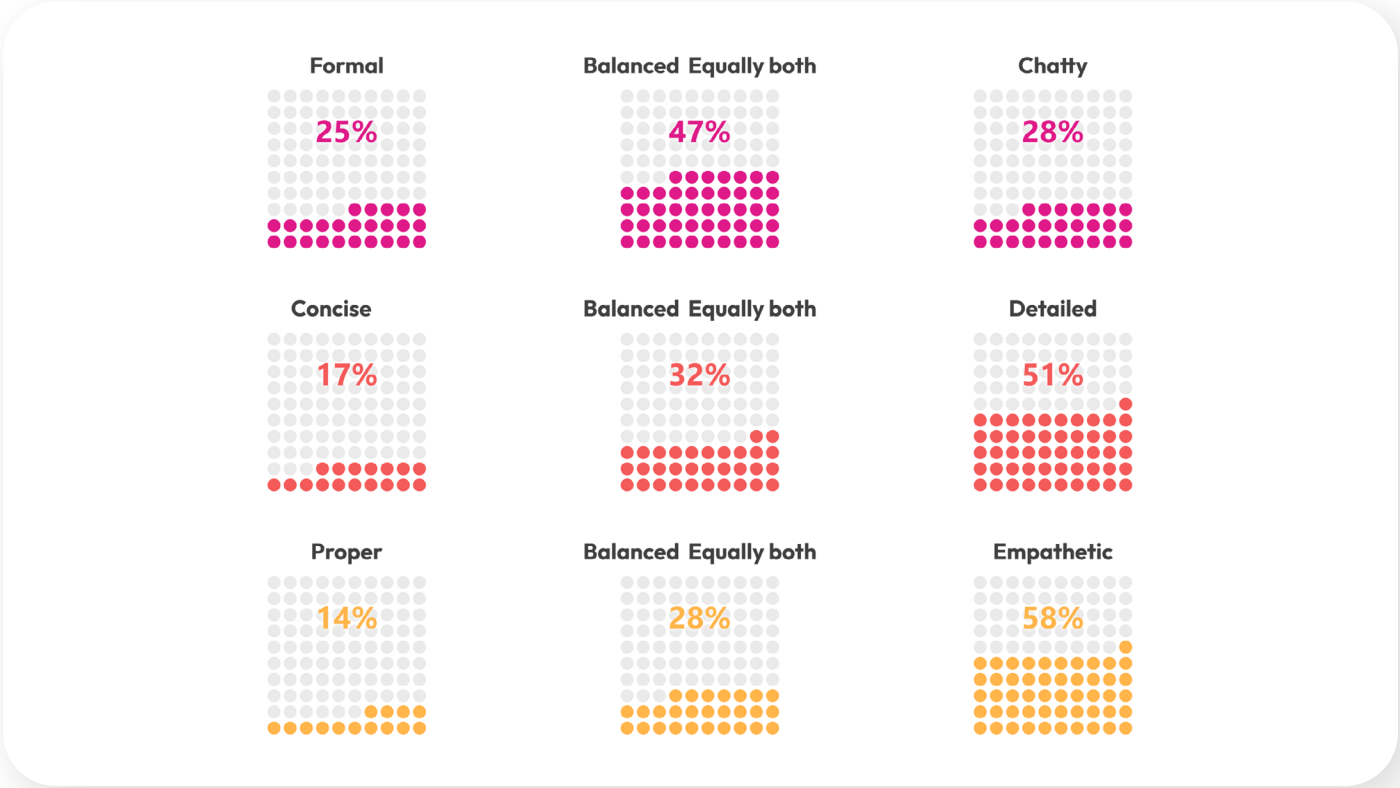

Figure 3 Selecting parameters for an agent's speaking style. You can choose whether you want the agent to speak slowly and in detail or quickly and to the point. Additionally, you can decide if you want the agent to be more empathetic, among other traits, and to personalize the agent's characteristics.

Hyper-personalization

Enterprise-grade agents must continuously learn from user interactions to refine how they respond. By identifying patterns in preferences, tone, and behavior, agents can anticipate needs and adjust their responses accordingly — whether that means offering a proactive solution or adapting their communication style. Customers increasingly expect agents to reflect their preferences — from tone and formality to visual style and even voice. Nearly half of consumers (49%) say they would like to customize their AI agent's characteristics (Amdocs, 2025 "Brand new frontier: the agentic era research"). Enterprise-grade agents should support this by dynamically adjusting their persona based on user profile, interaction type, and channel, to enhance the overall user experience, thereby providing personalized and relevant responses.

Enterprise-grade agents achieve this by learning from interactions and experiences over time. This kind of adaptive learning is supported by technologies like Amazon Bedrock, which enables fine-tuning and retrieval-augmented generation (RAG) pipelines, and NVIDIA NeMo™, which supports continues learning and multimodality RAG capabilities. Furthermore, NVIDIA NeMo™ can be deployed on Amazon Elastic

Compute Cloud P5 instances, powered by NVIDIA H100 GPUs, which accelerate AI deployment by reducing training costs by 4x and enabling faster solutions, enhancing hyper-personalized agent capabilities through scalable, efficient infrastructure.

Trust

Trust depends on the AI agent's ability to operate reliably, securely, and transparently. For enterprise-grade agents, trust is not a single feature — it's the result of consistent performance, responsible data handling, and adherence to safety and compliance standards. Protecting sensitive customer data is non-negotiable. Enterprise-grade agents must implement strong authentication, encryption, and access controls to safeguard Personally Identifiable Information (PII) and other confidential data. This includes securing both the interaction layer and the underlying data infrastructure.

AI agents must operate within clearly defined behavioral boundaries to prevent misuse, bias, or inappropriate responses. Guardrails help ensure that agents remain aligned with brand values, regulatory requirements, and ethical standards. NVIDIA's NeMo™ Guardrails framework provides a modular way to define and enforce these boundaries — including blocking unsafe topics, managing tone, and ensuring factual accuracy. This is particularly useful in customer-facing scenarios where agents must remain constructive and compliant even in emotionally charged or complex interactions.

AI agents must operate in accordance with evolving data protection laws, industry regulations, and internal governance policies. This includes global frameworks like GDPR, as well as sector-specific standards such as PCI-DSS for payment data. In telecommunications, compliance also extends to region-specific telecom regulations and lawful intercept requirements. Enterprise-grade agents must be auditable, policy-aware, and capable of enforcing data residency, retention, and access control rules.

Maintaining context in conversations is crucial for delivering accurate and relevant responses within the industry. It is essential to ensure that AI agents understand and retain the context of the interaction strictly within a business framework, avoiding unrelated topics, affinity, or profanity. For example, a customer should not be able to divert the conversation to non-business that goes outside of the context of the interaction, and the agent should maintain the conversation boundaries. Using this approach, the interaction optimizes token consumptionconsumption, and the agent can ensure that it responds appropriately to each query, building on the information already gathered.

Verticalized Agent Operations

Efficiently scaling AI agent deployments and ensuring future-proof operations in customer-facing enterprises, - particularly telcos, - requires the capabilities of scalability, deployment flexibility, LLMOps, and observability. AI agents must operate continuously and handle high volumes of concurrent interactions without performance degradation. Using multi-node GPU clusters, enterprises can parallelize model evaluation and fine-tune tasks, reducing iteration cycles to under an hour. This enables rapid experimentation and deployment of new models while maintaining consistent service levels.

Verticalized agents must be adaptable to diverse infrastructure environments. Enterprises benefit from deploying agents across on-premises, cloud, and hybrid environments, while remaining agnostic to the underlying LLMs, data lakes, and orchestration layers. NVIDIA NIM™ provides prebuilt, optimized inference microservices for rapidly deploying the latest AI models, offering an optimized model serving with a harmonized API that is compatible with both proprietary and GPT models. This allows organizations to integrate multiple model types without rearchitecting their systems. Additionally, Amazon Bedrock Marketplace now includes NVIDIA NIMs, offering integration with the AWS platform to improve performance, reduce latency, and enable adaptable deployment strategies.

Beyond infrastructure a structured LLMOps framework is essential for managing the lifecycle of large language models in production. This involves fine-tuning models on domain-specific data using multi-GPU setups, evaluating their performance across different use cases and customer-defined metrics, and automating model selection and upgrades based on benchmark results. Enterprises implement custom dashboards that monitor latency, accuracy, escalation rates, and multi-turn conversation quality. This observability supports continuous improvement and ensures that agents remain aligned with business goals.

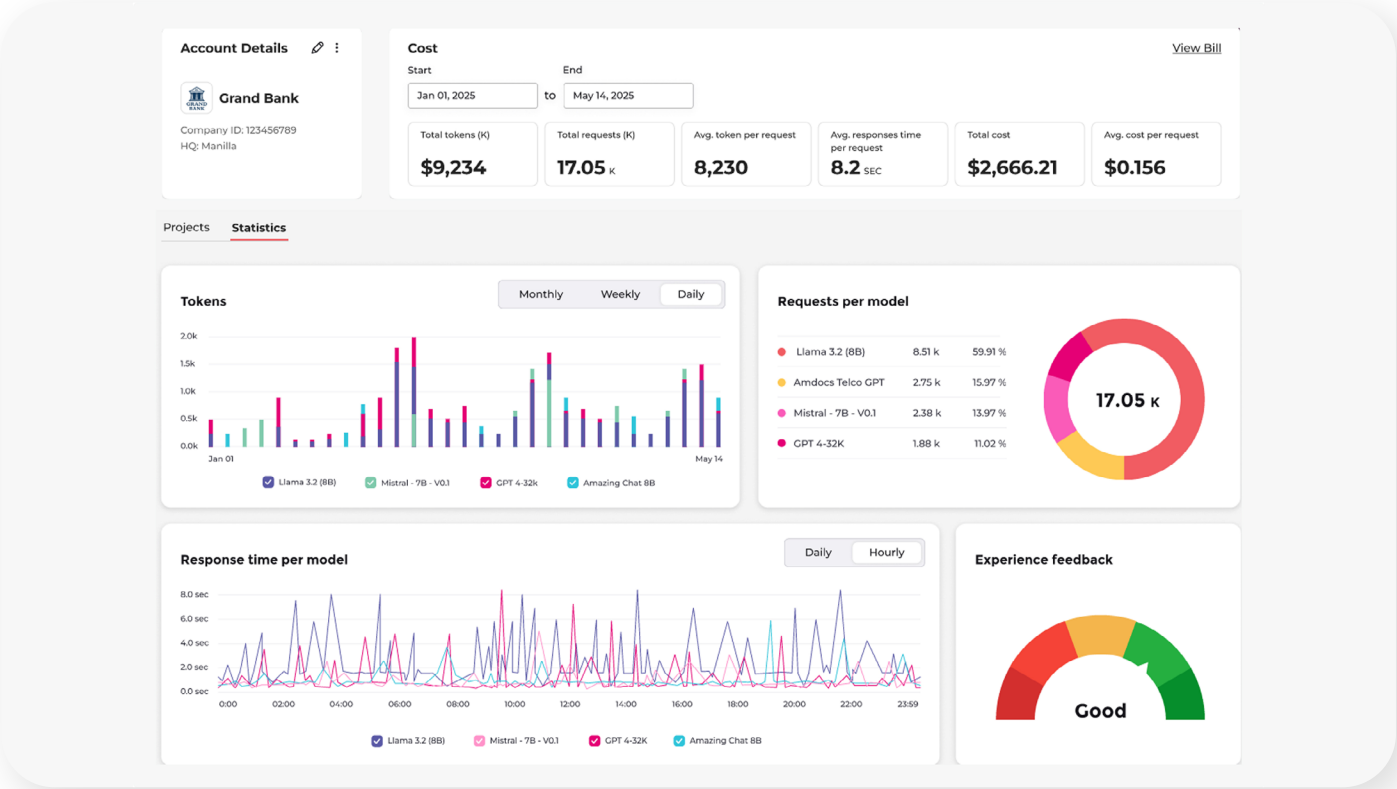

Figure 4 Example of monitoring and tracking dashboard for multiple agent models, associated costs and tokens

From Theory to Practice

To exemplify AI verticalization, this section examines practical implementations by AWS, NVIDIA, and Amdocs that enhance scalability, compliance, and operational efficiency. Key applications include proactive care, where agents predict billing anomalies to reduce support volume, and multi-agent collaboration, which streamlines customer interactions by coordinating specialized agents for tasks like billing, sales, and plan upgrades.

Telco care & sales agent

To demonstrate the real-world potential of verticalized, enterprise-grade AI agents, Amdocs, AWS, and NVIDIA partnered on the Telco GenAI Catalyst — a TM Forum initiative focused on transforming customer experience and operational efficiency in telecommunications through Generative AI.

The solution was built on a tightly integrated stack combining:

- **Amdocs amAIz:** A telco-grade GenAI platform that orchestrates multi agent systems, managing domain-specific skills, and enforcing compliance and operational standards.
- **AWS Cloud Services:** Delivers the scalable cloud infrastructure, including Amazon Elastic Cloud (EC2) with NVIDIA GPUs, Amazon DynamoDB and Amazon Simple Storage Service (S3), to support secure training, storage, and real-time inference.
- **NVIDIA AI Enterprise & DGX Cloud:** Powers the AI stack with NIM for optimized model deployment, Riva for speech services, NeMo™ Guardrails, and NVIDIA's Digital Humans for Customer Service

This architecture enabled seamless scalability, high availability, and rapid model iteration — all while maintaining enterprise-grade security and compliance.

Key use cases and capabilities

- **Proactive care:** Agents detected billing anomalies before invoices were issued, simulating upcoming charges and proactively notifying customers. This reduced inbound support volume and improved customer trust.
- **Multi-Agent collaboration:** The system supported domain-specific agents (e.g., billing, sales, plan upgrades) coordinated through a central router, enabling seamless handoffs and context sharing.
- **Conversational AI:** NVIDIA's Digital Humans for Customer Service and Riva enabled natural, multimodal interactions across voice and digital channels, enhancing accessibility and engagement.
- **Security and compliance:** The architecture incorporated enterprise-grade security controls, including encryption, access management, and auditability - aligned with telecom regulatory requirements.

Measurable outcomes

The catalyst delivered significant improvements across key performance indicators:

- **+60% reduction in token consumption**, lowering operational costs
- **Up to 30% improvement in response accuracy**, driven by domain-specific tuning.
- **Up to 80% reduction in query latency**, enabled by optimized inference pipelines.

These results demonstrate how a well-integrated GenAI stack — combining domain expertise, cloud-native infrastructure, and AI performance engineering — can deliver measurable business value while enhancing customer experience.

Figure 5 X-Ray behind the scenes of how the amAIz Agents Catalyst performed with NVIDIA and AWS, focusing on the process of understanding the prompt and connecting to the relevant agent. In this example, you can see that it selected the sales agent, who then utilized upsell skills, accessed the catalog, and offered a personalized package based on the customer's data and usage patterns

Autonomous networks acceleration with agentic AI and digital twins

To accelerate the journey toward autonomous networks (AN), Amdocs, AWS, and NVIDIA have partnered to deliver a transformative solution powered by agentic AI and digital twins - a strategic collaboration aimed at simplifying telecom operations and enhancing customer experience.

The solution is built on a robust, integrated architecture combining:

- **Amdocs Intelligent Networking Suite:** A service and network automation suite providing telco operations data and processes across inventory, assurance and orchestration.
- **Amdocs amAIz Suite:** A telco-specific GenAI platform that orchestrates multi-agent systems, leverages domain expertise, and ensures operational excellence.
- **AWS Cloud Infrastructure:** Provides scalable, secure services including AWS Glue, Amazon Neptune, Amazon Timestream for InfluxDB, SageMaker AI and Amazon Bedrock to support data ingestion, model training, real-time decision-making and autonomous agents
- **NVIDIA AI Enterprise:** Accelerates AI deployment with RAPIDS for data analytics, CUDA for AI and ML, NeMo for model customization, and NIM for optimized, faster and secured deployment.

This architecture enables organizations to build autonomous network agents that can handle complex queries, maintain conversation context, and leverage enterprise knowledge while scaling efficiently and maintaining security compliance.

Key use cases and capabilities

- **Self-service optimization:** Multi-agent collaboration between sales and network agents simplifies intent-based ordering, leveraging digital twins for real-time feasibility checks and seamless customer interactions.
- **Autonomous network operations:** AI-driven agents predict and prevent service-impacting faults using historical and real-time data, simulate resolutions using digital twins, and autonomously implement low-risk fixes

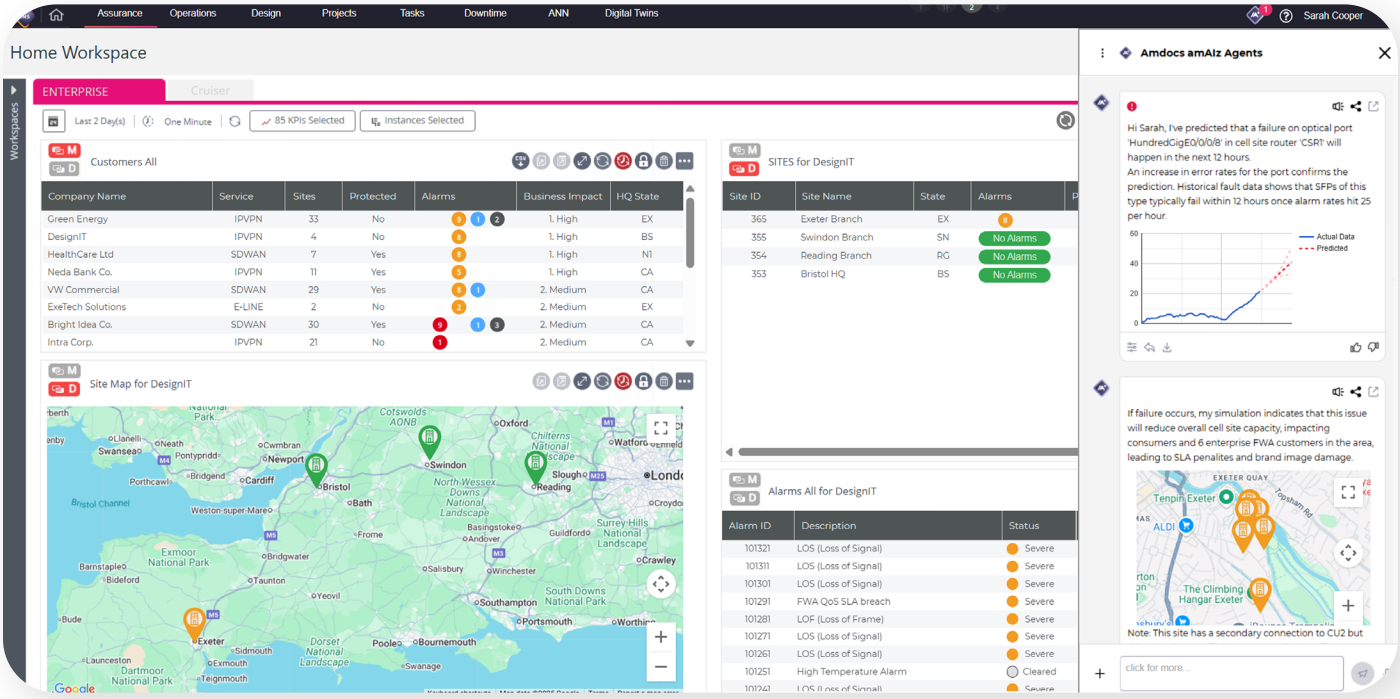

Figure 6 Amdocs amAIz Network Agent proactively notifying network operations engineer of predicted faults and its impact on customers

Business outcomes

The initiative lays the foundation for:

- Reduced Mean-Time-to-Repair (MTTR)
- Increased automated incident resolution
- Enhanced service quality and customer satisfaction, reducing churn

This collaboration illustrates how agentic AI and digital twins - when combined with deep telco expertise and cloud-native AI infrastructure - can drive meaningful progress toward fully autonomous networks.

Amdocs helps those who build the future to make it amazing. With our market-leading portfolio of software products and services, we unlock our customers' innovative potential, empowering them to provide next-generation communication and media experiences for both the individual end user and enterprise customers. Our employees around the globe are here to accelerate service providers' migration to the cloud, enable them to differentiate in the 5G era, and digitalize and automate their operations. Listed on the NASDAQ Global Select Market, Amdocs had revenue of $5.00 billion in fiscal 2024.

www.amdocs.com