ZTE TECHNOLOGIES

OCT 2025 | VOL. 27 • NO. 5 • ISSUE 220

Expert Views

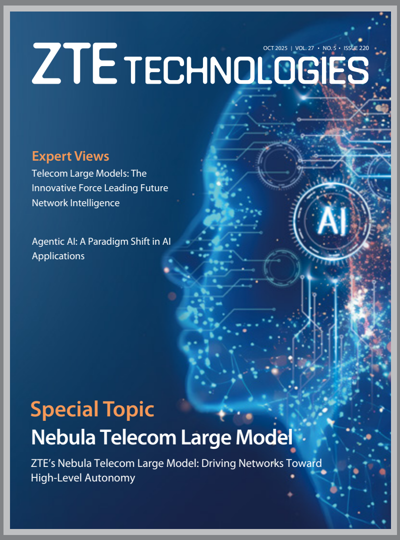
Telecom Large Models: The Innovative Force Leading Future Network Intelligence

Agentic AI: A Paradigm Shift in AI Applications

Special Topic

Nebula Telecom Large Model

ZTE's Nebula Telecom Large Model: Driving Networks Toward High-Level Autonomy


ZTE TECHNOLOGIES
OCT 2025 | VOL. 27 • NO. 5 • ISSUE 220
Expert Views
Telecom Large Models: The Innovative Force Leading Future Network Intelligence
Agentic AI: A Paradigm Shift in AI Applications
AI
Special Topic
Nebula Telecom Large Model
ZTE's Nebula Telecom Large Model: Driving Networks Toward High-Level Autonomy

ZTE TECHNOLOGIES

OCT 2025 | VOL. 27 • NO. 5 • ISSUE 220

Advisory Committee

Director: Liu Jian
Deputy Directors: Fang Hui, Sun Fangping, Yu Yifang, Zhang Wanchun
Advisors: Bai Gang, Dong Weijie, Hu Junjie, Hua Xinhai, Kan Jie, Li Weizheng, Liu Mingming, Lu Ping, Tang Xue, Wang Quan, Zheng Peng

Editorial Board

Director: Lin Xiaodong
Deputy Director: Lu Dan
Members: Deng Zhifeng, Dai Yanbin, Huang Xinming, Jiang Yonghu, Kong Jianhua, Liang Dapeng, Liu Shuang, Lin Xiaodong, Lu Dan, Ma Xiaosong, Shi Jun, Yu Fanghong, Yang Zhaojiang

Sponsor: ZTE Corporation

Published by ZTE Technologies Editorial Office
General Editor: Lin Xiaodong
Deputy General Editor: Lu Dan
Editor-in-Chief: Liu Yang
Circulation Manager: Wang Pingping

Editorial Office Address: No. 55, Hi-tech Road South, Shenzhen, P.R. China
Circulation Office Address: 12F Kaixuan Building, 329 Jinzhai Road, Heifei, P.R. China
Website: www.zte.com.cn/global/about/publications
Email: magazine@zte.com.cn

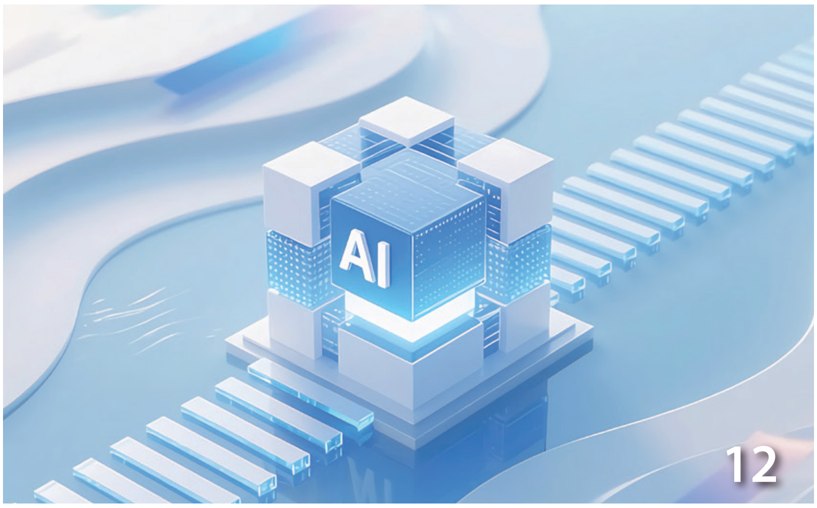
Statement: This magazine is a free publication for you. If you do not want to receive it in the future, you can send the "TD unsubscribe" mail to magazine@zte.com.cn. We will not send you this magazine again after receiving your email. Thank you for your support.

CONTENTS

Foreword

Expert Views

Special Topic: Nebula Telecom Large Model



Success Stories

Zhang Wanchun

SVP of ZTE

Nebula Telecom Large Model

Opens a New Chapter in High-Level Autonomous Networks

The new wave of technological revolution and industrial transformation is reshaping the global innovation landscape at an unprecedented pace. The rise of AI, particularly large language models (LLMs), is driving society toward an era of AI democratization. In the field of communications, AI large models have become a core technological driver, enabling the evolution of Autonomous Networks to Level 4 and beyond.

ZTE has been advancing the application of AI technology in operators' O&M scenarios, driving a paradigm shift in AI. Through innovations in AI large models, agents, and digital twins, ZTE helps operators optimize and automate processes, significantly enhancing the automation and intelligence of network O&M and facilitating the evolution toward "Agentic Operation."

Leveraging its self-developed Nebula Telecom Large Model, ZTE applies domain-specific knowledge for model pre-training and reinforcement learning, while integrating technologies such as knowledge graphs, structured data AI models, and digital twins to effectively mitigate the hallucination issues commonly associated with LLMs. Simultaneously, the knowledge flywheel mechanism enables online iteration and scenario-based generalization, bridging the gap between general-purpose LLMs and domain-specific requirements, and gradually turning them into new productive forces for operators.

At the beginning of this year, ZTE upgraded its autonomous network solution, AIR Net (also known as AI Reshaped Network), built on the Nebula Telecom Large Model, big data, and digital twin engines, to drive the intelligent transformation of infrastructure and both single-domain and cross-domain operations. Through open decoupling, ZTE helps operators transition from fragmented AI implementations to fully integrated, end-to-end intelligent systems. Based on this, ZTE focuses on high-value scenarios for operators, establishing a value assessment system and relevant case studies to achieve measurable business outcomes such as improved service quality, increased revenue, cost reduction, and efficiency improvements, ultimately realizing a business closed loop.

Looking ahead, ZTE will continue to build on communications-centric large models, deepen collaboration with global industry partners, jointly advance AI large model technologies, explore new value propositions, foster an open and mutually beneficial ecosystem, and contribute to the digital and intelligent transformation of operators' networks. ZTE TECHNOLOGIES

Telecom Large Models:
The Innovative Force Leading Future Network Intelligence

Kang Honghui

Chief Architect for Wireless
Network Intelligence, ZTE

Liu Kunlin

Chief Engineer of Nebula
Telecom Large Model R&D, ZTE

The deep integration of AI with next-generation information technologies has become a key driver of autonomous network transformation. With the widespread application of large models in the networking field, telecom large models are reshaping network operation and service innovation, driving the intelligent evolution of networks.

Proliferation of Telecom Large Models and Their Applications in the Large Model Era

Since 2023, large language models (LLMs), represented by ChatGPT, have ignited a global technological race. In early 2025, DeepSeek released its open-source model DeepSeek-R1, which rapidly gained traction across verticals such as finance, healthcare, and smart manufacturing, due to its strong reasoning capabilities and cost efficiency. In the telecom industry, the rise of new businesses—including the low-altitude economy, the Internet of Things (IoT), vehicle-to-everything (V2X), and immersive extended reality (XR)—as well as operators' demands for digital-intelligent transformation, has accelerated the adoption of generative AI (GenAI). This has fueled rapid growth in autonomous networks and intelligent agent applications built on telecom large models.

The rapid development of telecom large models has been driven by active investments from both operators and equipment vendors. Major Chinese operators are leveraging their data advantages to accelerate the deployment of industry-specific large models. China Mobile has released its self-developed AI platform, "JiuTian", which accumulates 450 AI capabilities. China Telecom's Xingchen large language, speech and multimodal models have completed “dual recordation” of algorithms and services. China Unicom's Yuanjing large model has developed over 40 industry-specific models. According to IDC, the market size of large models in China's telecom industry grew by 67% year-on-year in 2024, with multimodal and scientific computing models emerging as new growth points.

On the equipment vendor side, ZTE is pursuing a dual-track approach of "full-stack in-house R&D + ecosystem collaboration." Its Nebula Telecom Large Model, available in versions ranging from 7B/14B to 100B parameters, supports scenarios such as network operation and maintenance, fraud detection, and signaling analysis through an architecture of "Nebula large model + agent factory + serialized applications".

Additionally, the open-source ecosystem fosters technological democratization and advances the development of telecom large models—DeepSeek's open-source strategy has significantly lowered industry entry barriers. By Q1 2025, over 60% of telecom enterprises in China developed customized solutions based on open-source models. This "open sharing + vertical specialization" model is reshaping the innovation path of intelligent network technology.

Telecom Large Models Reshape Autonomous Network Evolution

Large Models Reconstruct the Intelligent Engine of Autonomous Networks

Autonomous networks achieve automation and intelligence through a "three-layer, four-closed-loop" architecture (Fig. 1). To advance toward L4 and L5 autonomous networks, TM Forum has outlined key enabling technologies, including network AI large models, trustworthiness technology, and digital twins. These technologies will support the AI requirements embedded in 3GPP networks, driving autonomous networks toward higher levels of intelligence. In this evolution, telecom large models serve as the intelligent engine. Together with the data and digital twin engines, they form the digital-intelligent technology foundation of autonomous networks.

Large Models Reconstruct the Technical Architecture of Autonomous Networks

Telecom large models are redefining the technical architecture of autonomous network evolution into a three-tier system comprising "full-scenario AI paradigms + digital-intelligent engines + intelligent agents." This architecture introduces serialized models layer by layer, facilitating embedded AI within networks and full-scenario AI integration. For example, in the transition from single-domain closed loops to cross-domain collaboration, operators can begin by achieving closed loops in single-domain optimization scenarios—such as wireless network optimization—by leveraging the intelligent analysis capabilities of large models. This can then be expanded to end-to-end, cross-domain scenarios for

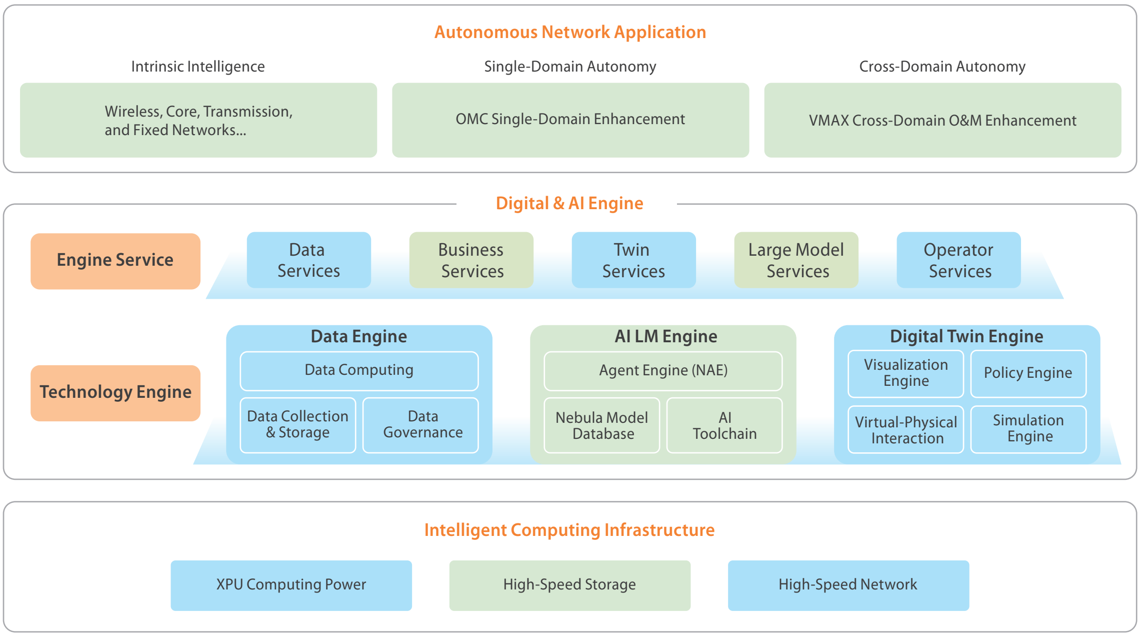


Fig. 1 Technological architecture of autonomous networks.

global network management, characterized by "traceable history, visible reality, and predictable future." Enabled by the integrated "perception–decision–execution" capabilities of telecom large models, successful applications have been demonstrated in intent understanding, autonomous learning, long-process closed-loop optimization, and accuracy improvement.

For instance:

- **In planning:** For 5G-A network construction, digital twin tools assisted by large models can simulate over 100,000 channel combinations, shortening planning cycles.
- **In operation and maintenance:** China Mobile's intelligent O&M system, based on DeepSeek, has significantly improved fault localization efficiency, paving the way for minute-level fault response.
- **In operations:** ZTE's fraud detection large model analyzes communication behavior patterns to identify new types of telecom fraud, greatly reducing false positives.

Large Models Reconstruct the Agent Paradigm of Autonomous Networks

Telecom large models have given rise to a new generation of intelligent agents for networks, shifting traditional autonomous network O&M toward an agent-based paradigm. With capabilities in autonomous learning, task execution, and multi-task collaboration, these intelligent agents provide a new pathway for telecom operators' intelligent transformation. By deploying agents, operators can achieve highly automated network management and optimization, enhancing self-perception and self-healing capabilities, thereby significantly reducing operational costs and improving network reliability.

The release of OpenAI's o1 and DeepSeek's R1 models—both excelling in complex reasoning tasks—has inaugurated a new paradigm of "slow-thinking" large models, offering innovative solutions for network O&M and intelligent agents. These models, along with their reasoning-enhancement technologies, enable task orchestration, problem analysis, flexible decision-making, and optimized execution in complex business scenarios. Meanwhile, agent applications are shaping a full-scenario AI paradigm across areas like interaction methods, interface designs, product architecture, business capabilities, and development and delivery. This agent paradigm accelerates the path toward achieving L4 autonomous networks.

Telecom Large Models Evolve Toward Network World Models

The future evolution of telecom large models will advance toward stronger performance, lower resource consumption, and diversified interaction modes, while also progressing toward network world models that are deeply integrated with network services.

Innovation Directions for Telecom Large Models

Innovation in telecom large models is advancing along several key directions: model architecture innovation, multimodal fusion, domain-augmented reinforcement learning, and the further convergent evolution of models and agents.

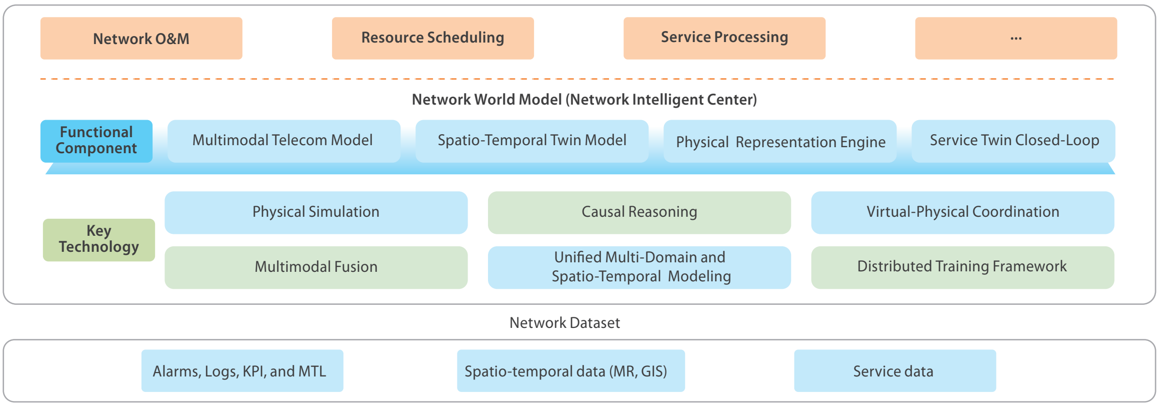
- **Model Architecture Innovation**

In telecom large model scenarios that prioritize localized deployment, architectural innovation is essential to achieve higher performance and smaller model sizes. It is worth considering the introduction of a spatio-temporal decoupled routing mechanism on top of the traditional Mixture-of-Experts (MoE) architecture, integrating temporal sequence models with spatial data to construct a spatio-temporal network model, which can then be further fused with language models.

In terms of coordination mechanisms, a bidirectional knowledge distillation approach can be employed between central large models and edge small models. The central model extracts global features (e.g., nationwide network traffic patterns), while edge models capture local characteristics (e.g., regional base station deployment differences), enabling collaborative model evolution through a federated learning framework with dynamic weight exchange.

- **Multimodal Fusion**

Future large models will integrate multimodal data—including text, speech, images, network signaling, alarms, and logs. This fusion will enhance the accuracy of network state perception and enable more immersive interaction methods and personalized



Network O&M
Resource Scheduling
Service Processing
...
Network World Model (Network Intelligent Center)
Functional Component
Multimodal Telecom Model
Spatio-Temporal Twin Model
Physical Representation Engine
Service Twin Closed-Loop
Key Technology
Physical Simulation
Causal Reasoning
Virtual-Physical Coordination
Multimodal Fusion
Unified Multi-Domain and Spatio-Temporal Modeling
Distributed Training Framework
Network Dataset
Alarms, Logs, KPI, and MTL
Spatio-temporal data (MR, GIS)
Service data

Fig. 2 Network world model.

customer service solutions, ultimately improving the user experiences.

- **Lightweighting and Edge Deployment**

  Large parameter models can be compressed using quantization and distillation to support low-power inference at edge nodes, meeting the real-time requirements of industrial IoT.

- **Domain-Specific Reinforcement Learning**

  Enhance large models' expertise in vertical scenarios by incorporating telecom-specific knowledge graphs, such as protocol stack rules and fault case libraries.

Co-Evolution of Digital Twins and Large Models

The co-evolution of large models and digital twin technology is a form of bidirectional empowerment. Digital twins provide telecom large models with virtual-physical mapping testbeds and synthetic data, supporting twin-driven model training. In turn, large models enhance the simulation and reasoning capabilities of digital twins, equipping them with dynamic optimization abilities. While digital twins enable real-time mapping of physical network states, large models can generate optimal resource scheduling strategies through reinforcement learning. This enables a "digital twin–model training–network deployment" closed loop for network functions, transforming networks from passive monitoring to proactive design.

Toward a Unified Network World Model

In the 6G era, telecom large models will evolve into "network world models," enabling unified representation and prediction of physical networks, business demands, and user behaviors, and serving as the intelligent core of future networks (Fig. 2).

Key features include:

- **Multi-domain unified modeling:** Integrate wireless, core, transport, and application-layer data with spatio-temporal digital twin models for global modeling.
- **Causal reasoning capabilities:** Shift from correlation analysis to causal inference.
- **Autonomous evolution mechanisms:** Enable the self-iterating evolution of models through continuous learning and feedback loops, allowing adaptation to the dynamic demands of new services.
- **Twin closed-loop:** Future intelligent agent networks will possess end-to-end digital twin closed-loop capabilities in various scenarios.

Outlook

Telecom large models are evolving from technical tools into the core engines of network intelligence, reshaping the technological foundation of autonomous networks. They are driving a shift from single-domain optimization to full-domain collaboration, and from manual intervention to autonomous decision-making.

With the advancement of 6G network intelligence architectures, the integration of network world models—combining digital twins, edge inference, and causal learning—will propel autonomous networks toward a qualitative leap, from "functional autonomy" to "cognitive intelligence."

This intelligence revolution, powered by telecom large models, is transforming the technological form of network infrastructure and redefining the value boundaries of autonomous networks. ZTE TECHNOLOGIES

Agentic AI: A Paradigm Shift in AI Applications

Gao Yanqin

Chief Planning Engineer for Telecom Large Model and RAN Agents, ZTE

Du Yongsheng

Chief Technology Engineer for Telecom Large Model and RAN Agents, ZTE

We are entering an era of rapid AI expansion, essentially a process of iterative technological paradigm shifts that continually open new application scenarios. Today, large model-powered AI agent technology is ushering in a profound paradigm change—from tool-based AI to Agentic AI (autonomous agents). With a closed-loop of planning, perception, decision-making, and execution, Agentic AI is transforming autonomous networks from manually driven to agent-led systems.

AI Paradigm Shift: From Tool Execution to Autonomous Execution

AI agent applications are evolving from tool orchestration and execution to agent systems that can act on behalf of users—undertaking complex tasks, making decisions, and adapting to ever-changing environments.

Large Models Reshaping the Cognitive Foundation of Agents

Large language models (LLMs), pre-trained on vast amounts of data, establish a general world knowledge base. They can understand complex contexts, tackle cross-domain tasks, and demonstrate rudimentary reasoning and creative capabilities. As such, they form the cognitive foundation of agents, which can be summarized in three dimensions:

- **Universal intelligent base:** The LLMs act as the agent's "brain," providing understanding, reasoning, and generation abilities.
- **Environment perception and action:** Agents actively interact with their environment through APIs, sensors, and tool interfaces.
- **Goal-driven and evolutionary:** Based on preset goals, agents plan execution paths, dynamically adjust strategies, and continuously optimize through feedback.

In this paradigm, AI shifts from being merely a "tool" to becoming an intelligent agent capable of independent decision-making, long-horizon operation, and dynamic evolution. Its operational logic follows a task-goal loop: the model is invoked to make decisions, tools execute accordingly, and results are fed back to the model until termination.

LLM-based agents can autonomously decompose tasks, invoke toolchains, and dynamically adjust strategies, creating a closed-loop decision-making system. This "goal-driven, autonomous planning" model marks a fundamental shift in AI applications—from functional modularization to agent-centricity. Gartner predicts that by 2026, more than 80% of enterprises will deploy AI agents to restructure business processes, achieving efficiency gains of 40% to 60%.

AI Agents Powering a New Model of AI Production

Current LLM providers, such as OpenAI, Anthropic, Alibaba, and ByteDance, are transitioning from providing single API outputs to building agent ecosystem platforms. Microsoft, for example, has launched Copilot Studio, allowing enterprises to customize exclusive agents and integrate internal data and business processes. ByteDance has rolled out Cozi and Cozi Space as platforms for internet-based AI agents and enterprise applications.

The widespread adoption of AI agents is fueling the growth of rapid AI application development platforms, which are becoming central hubs for AI value creation. The new model of AI production centers on interactive AI agent design, end-to-end rapid development, delivery, deployment, operation, and integration within the AI ecosystem (see Fig. 1).

Agentic AI Evolution: From Passive Response to Active Adaption

AI is entering a new agentic stage, evolving from passive instruction execution to active learning, adaptation, and optimization. Through self-iteration and interaction with the environment, such systems can dynamically adjust their strategies to achieve goal-driven, continuous evolution.

As defined by Anthropic, LLM-based agents can autonomously direct their own actions and tool usage, maintaining control over how tasks are accomplished. Although still in a transitional phase, Agentic AI has seen milestones—such as the release of Manus (March 6), the launch of AI Superframe by Quark (March 13), the introduction of Agent TARS by ByteDance, and the booming popularity of Genspark. These developments signal a shift from "tool executors" to highly intelligent entities capable of goal-oriented collaboration and active evolution.

Agentic AI Working Mode: Target-Driven Multi-Agent Coordination

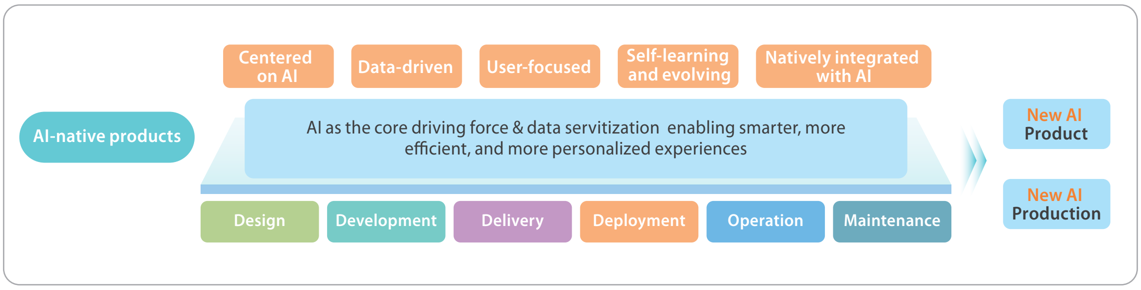
To understand the revolutionary changes brought about by the working mode of Agentic AI, it is useful to compare it with traditional workflows.

Take the search scenario as an example—this transformation is particularly evident—as it not only reconstructs the technical implementation path but also redefines the underlying logic of human-machine collaboration. A typical Agentic AI search process might proceed as follows:

- A user raises a query, and the agent analyzes and decomposes it to infer true intent.
- If the query is ambiguous, the agent proactively seeks clarification from the user.
- It then selects the appropriate search method—general-purpose search or specialized data sources depending on context.
- Finally, it integrates the search results and output them in a way aligned with the user's intent.

The LLM continuously learns from search

Fig. 1 Framework of AI-native product production.


Centered on AI
Data-driven
User-focused
Self-learning and evolving
Natively integrated with AI
AI-native products
AI as the core driving force & data servitization enabling smarter, more efficient, and more personalized experiences
Design
Development
Delivery
Deployment
Operation
Maintenance
New AI Product
New AI Production

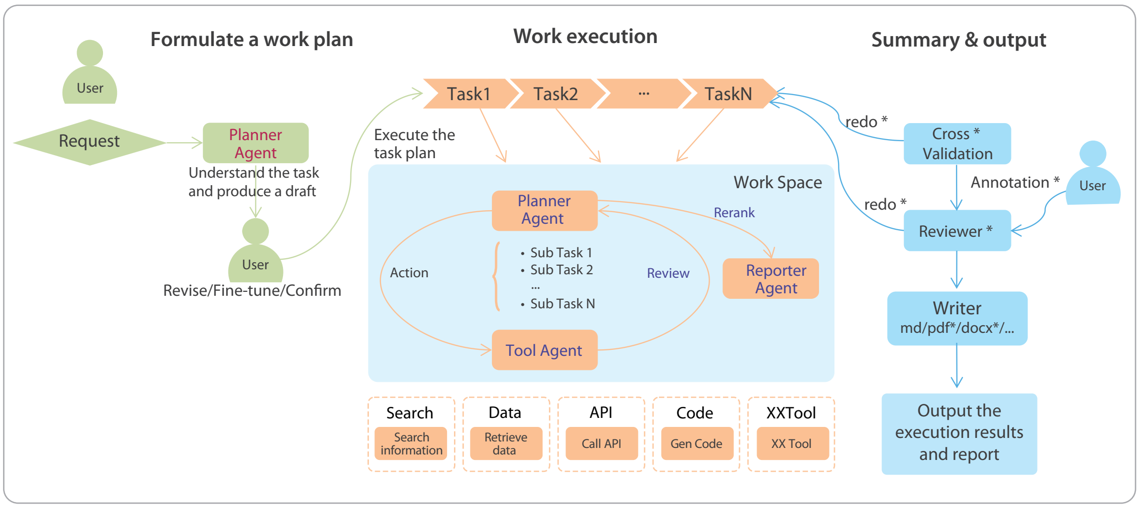


Fig. 2 Definition of the general working mode of Agentic AI.

processes and history, enabling the agent to autonomously determine search directions and strategies. Each decision and inference is clearly logged, achieving a degree of interpretability.

This case demonstrates that, compared with traditional rule-based or workflow-concatenation approaches, Agentic AI shifts from static workflows to dynamic decision-making. As shown in Fig. 2, tasks are completed through the collaboration of multiple agents within a full-stack closed loop comprising planning, decision-making, execution, and evaluation.

Agentic AI has achieved breakthroughs in three key dimensions:

- **Goal-driven dynamic planning:** Free from preset workflows, it decomposes tasks and adjusts strategies based on real-time environments.
- **Closed-loop self-learning and optimization:** It improves decision-making paths through continuous feedback and learning, without manual tuning.
- **Intelligent resource integration:** It proactively invokes heterogeneous systems (e.g. APIs, databases, toolchains) instead of passively awaiting instructions.

Application Scenarios of Agentic AI

We extend the Agentic AI paradigm to autonomous networks, where its impact is equally significant. In wireless access networks, with fault handling and network optimization as two examples, we illustrate a closed-loop system under the Agentic AI paradigm, encompassing perception, decision-making, execution, evolution, enabled by data sharing, strategy linkage, and task relay.

Troubleshooting Coordination: Automated Closed-Loop Work Orders

In the event of a sudden service interruption at a gNodeB, the system automatically executes the following steps:

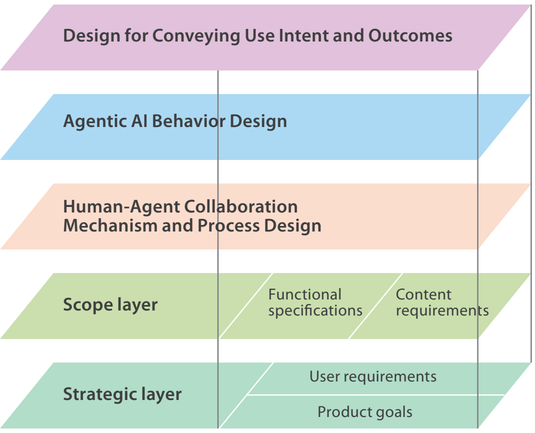
- **Fault perception:** The fault-management agent detects in real-time that the RRC connection success rate has dropped from 99% to 72%, accompanied by a 6 dB decrease in the average signal-to-interference-plus-noise-ratio (SINR), triggering a Level-3 alarm.
- **Root cause diagnosis:** The agent queries the knowledge graph, matches the historical case database, and invokes APIs, commands, and tools for data inspection and root cause diagnosis, ruling out hardware faults (equipment health indicators are normal). It primarily provides intelligent resource integration.
- **Fuzzy analysis:** Based on the topology spanning the base station, transmission, power and environment monitoring, and the core network, the agent identifies co-channel interference caused by deviations in antenna downtilt of adjacent cells, showcasing goal-driven dynamic programming.

- **Automatic repair:** The agent provides repair plans, executes automated actions, and prompts users to confirm high-risk operations that may affect services.
- **Dynamic parameter adjustment:** The agent executes a command to correct the antenna downtilt angle of the faulty base station from 8° to 12°, reducing coverage range.
- **Result feedback:** The agent checks the outcomes of parameter adjustments and initiates indicator monitoring.
- **Indicator verification:** The connection success rate rebounds to 98%, and the SINR returns to 18 dB.
- **Automatic work order report generation:** The report is automatically delivered to O&M personnel through the mobile terminal and marked as “No manual intervention required.”
- **Self-learning optimization:** The results are stored in the agent's shared memory. Using reinforcement learning, the agent extracts scenarios, rules, and major diagnoses and localization actions from the repair process for future cases, enabling closed-loop self-learning.

Optimization Coordination: Predictive Network Optimization

During a large-scale event, the network is affected by a surge in traffic. The system responds as follows:

- **Joint prediction (24 hours before the event):** The optimization agent predicts a 10-fold traffic peak around the venue, while the fault management agent pre-checks and confirms that the load margin of nearby base stations is insufficient.



Design for Conveying Use Intent and Outcomes
Agentic AI Behavior Design
Human-Agent Collaboration Mechanism and Process Design
Scope layer
Functional specifications
Content requirements
Strategic layer
User requirements
Product goals

Fig. 3 Agent design elements.

- **Dynamic pre-configuration (two hours before the event):** Agents mainly provide intelligent resource integration. The optimization agent activates dormant cells to expand network density, while adjusting the QoS policy to prioritize video bandwidth. The fault-management agent simultaneously initiates a system health check and performs stress testing on the capacity-expanded base stations to rule out hardware risks.
- **Network optimization (during the event):** When the instantaneous number of users exceeds the forecast by 15%, the optimization and fault-management agents respond collaboratively. The optimization agent initiates spectrum offloading, while the fault-management agent monitors CPU temperatures and dynamically restricts non-urgent services to ensure system stability. At this stage, agents mainly demonstrate goal-driven dynamic programming capabilities.
- **Post-event notification:** The agent automatically generates a network quality report and pushes it to users via SMS or app, stating "The average user rate during the peak period was 52 Mbps, with a compliance rate of 97%".

Design and Technologies of Agentic AI

Agentic AI Design Paradigms

The design of Agentic AI is evolving towards user-adaptive experiences, focusing on three key levels, as shown in Fig. 3.

- **Conveying user intent and outcomes:** Dynamically map user intent and system responses through interface elements to provide a visual closed-loop of operation paths and feedback.
- **Agentic AI behavior design:** Incorporate anthropomorphic feedback and expectation management to demonstrate the interpretability, controllability, and human-like behavior patterns of AI agents.
- **Human-AI collaboration mechanisms and processes:** Focus on bidirectional adaptability between humans and AI by building an evolvable collaboration framework with dynamic role allocation, context awareness, transparent

decision-making paths, and progressive trust building. This enables AI agents to proactively predict needs, explain behaviors, and accept interventions, ultimately supporting seamless task collaboration.

Core Technologies of LLM-Based Agents

The more an agentic system relies on LLMs for behavioral patterns, the higher its autonomy. LLM-based Agentic AI is centered on four core modules: planning, memory, tools, and perception.

Key technologies of the planning module include:

- **Hierarchical task reasoning:** Perform global planning based on LLMs, integrating mind maps with reinforcement learning to achieve multi-level goal decomposition and dynamic path optimization.
- **Causal reasoning enhancement:** Integrate structural causal relationships with empirical data to predict the long-term impact of actions and avoid short-sighted decisions.
- **Multi-agent collaboration:** Leverage techniques such as real-time multi-agent collaboration, shared memory, and multi-agent confrontation to support task allocation and conflict resolution.

Key Technologies of the memory module include:

- **Long-term memory compression:** Store key historical information and realize efficient retrieval and association via the self-attention mechanism.
- **Short-term memory management:** Cache dialogue states and environmental contexts to maintain coherence in multi-turn interactions.
- **Knowledge distillation and update:** Dynamically expand external knowledge bases through continual learning or retrieval-augmented generation (RAG) to avoid model hallucinations.
- **Dynamic enhancement of memory:** Construct a dynamic memory bank to store interaction trajectories and apply the attention mechanism to extract key experiences, enhancing memory dynamically.

Key technologies of the tool module include:

- **Automatic tool discovery and usage:** Automatically construct tool calling methods based on semantic embedding matching of tool description texts. As the real world is open and dynamic with new tools, APIs, and data sources continuously emerging, an intelligent system must generalize like humans by understanding functional descriptions of new tools and incorporating them into its own capability set.
- **Secure sandbox verification:** Pre-execute high-risk operations (e.g., network requests) in a restricted environment (e.g., a Docker container), and return results to the main process only after verification.

Key technologies of the perception module include:

- **Multimodal information fusion:** Utilize cross-modal representation alignment to uniformly process text, image, and speech inputs to construct environmental state representations.
- **Dynamic environment modeling:** Predict environmental changes through a world model to help agents anticipate the impact of actions.
- **Active perception and attention control:** Optimize perceptual focus with reinforcement learning, prioritizing high-value information (e.g., user preferences and habits in conversations).

The Evolutionary Outlook for Agentic AI

AI agents and LLMs are advancing in a mutually reinforcing manner and evolving at a fast pace. In certain domains, they have already approached or even surpassed human experts. We must innovate boldly, while allowing time for the technology to be validated in practice.

Technically, AI is progressing from perceptual intelligence to autonomous decision-making, and multimodal integration will enable agents to interact with the physical world. In application, deployment will begin in finance and medical care before expanding to other vertical fields. Socially, human-AI collaboration will reconstruct workflows and drive new ethics and value frameworks. The design approach is shifting from a "Reason + Tool" model to a "Learning + Reason" model.

Ultimately, Agentic AI will bring a comprehensive upgrade to human cognition, collaboration systems, and values—ushering in a new chapter in human development. ZTE TECHNOLOGIES

ZTE's Nebula Telecom Large Model

Driving Networks Toward High-Level Autonomy

Zheng Peng

General Manager of Data Intelligence and Service Products, ZTE

Wang Chengchun

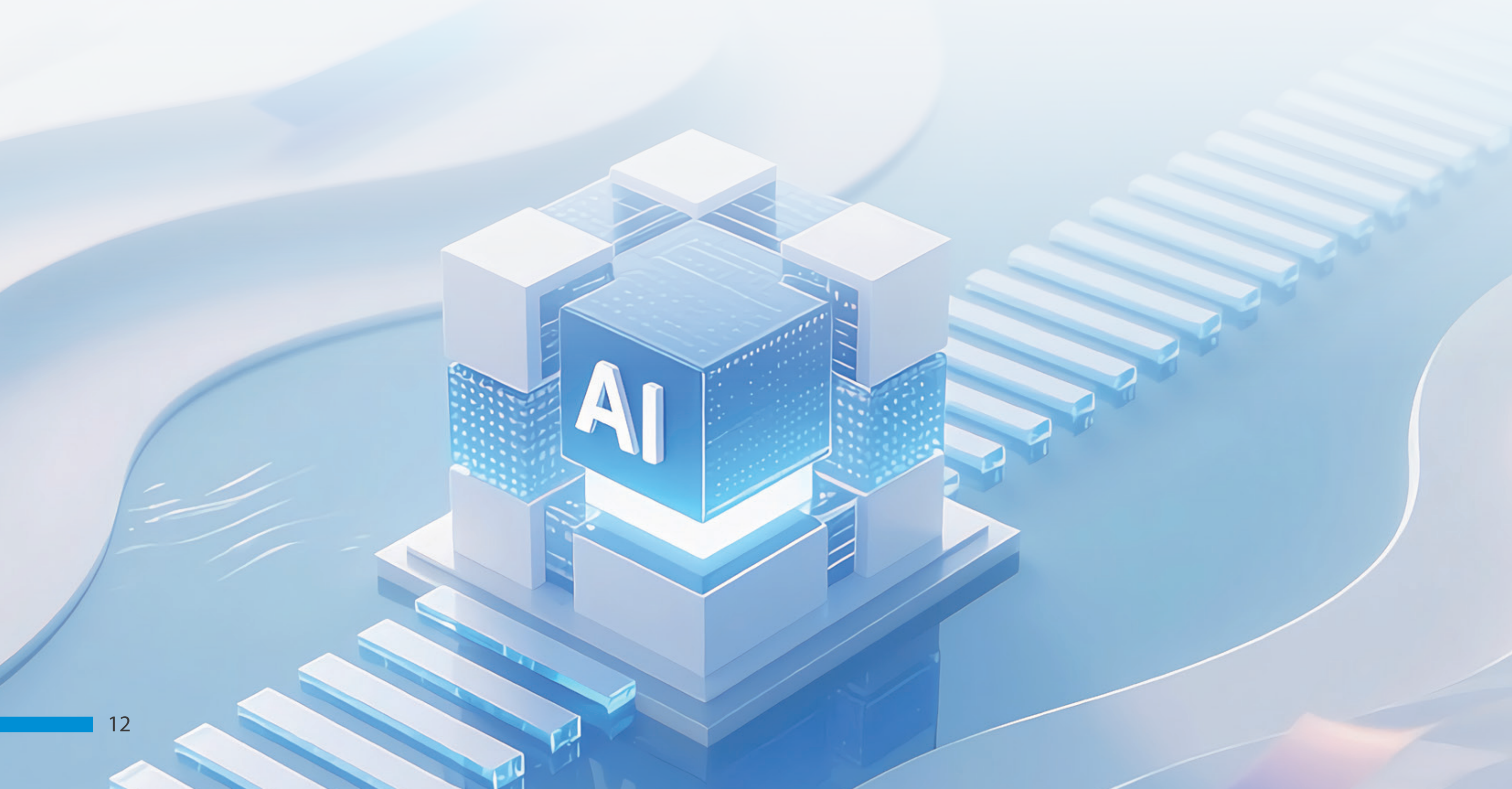
Chief Planner of Telecom Large Model, ZTE

Fan Xuefeng

Chief Engineer of Big Data Products, ZTE

Today's communication networks face two major challenges: rapidly increasing complexity and the growing demand for intelligence. With 5G-A commercialization and the advent of 6G, ubiquitous connectivity is driving emerging industries such as XR and Industrial Internet, accelerating business value creation. Meanwhile, the integration of new technologies such as AI large models is reshaping networks into a new intelligent paradigm.

ZTE's AIR Net high-level autonomous network solution provides full-scenario, closed-loop automation across multiple services and domains throughout the entire lifecycle with an open,



decoupled design. By leveraging AI large models, intelligent agents, and digital twins, it accelerates the path to high-level autonomous intelligence, guided by three key values:

- **Commercial effectiveness:** Enabling end-to-end closed loops in high-value autonomous network scenarios, reducing labor, costs, and time, while supporting machine decision-making.
- **O&M efficiency:** Shifting from manual-driven to data-driven intelligence, boosting fault-handling efficiency severalfold.
- **Ecosystem value:** Driving collaboration between vendors and operators in vertical industries to meet diverse demands, creating a new "network as a service (NaaS)" ecosystem.

Building a Flexible, Dynamic, and Rigorous Intelligent Foundation

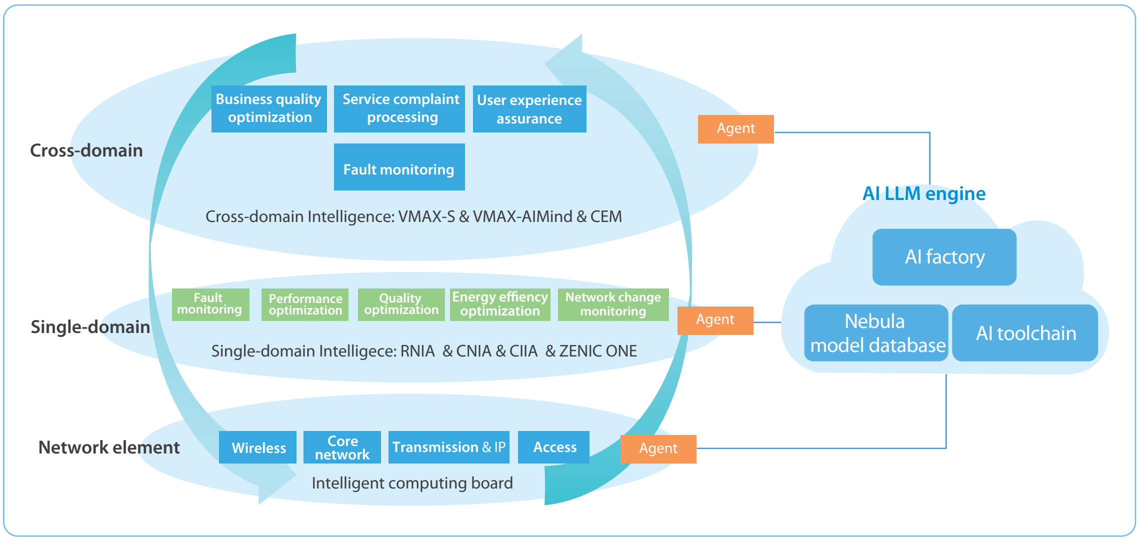
The core goal of ZTE's AIR Net autonomous network architecture is to enhance autonomous operation and optimize the closed-loop efficiency of business responses, eliminating process gaps and bottlenecks caused by manual intervention. Current networks, while capable of basic automation, such as alarm filtering and simple work order dispatch, still relies largely on predefined rules and local AI models. This approach has clear limitations in complex scenarios such as cross-domain fault localization and quality degradation analysis. To truly leap forward, the network requires an "intelligent hub" with deep network cognition—an intelligent decision-making engine powered by large models. By tightly integrating general intelligence with domain-specific knowledge, the network will transition from automation to intelligence.

ZTE’s Nebula Telecom Large Model (see Fig. 1) addresses this with breakthroughs in four dimensions—foundation flexibility, task autonomy, knowledge dynamics, and reasoning rigor—enabling the network to evolve from rule-based automation to cognitive automation.

Full-Stack Large Model System: Seamless Engine Switching for "Foundation Flexibility"

The complexity of network environments lies in hardware heterogeneity, platform diversity, and long-tail scenarios. Through a layered, decoupled architecture, the Nebula Telecom Large Model realizes seamless engine switching, allowing the foundation model to be replaced without modifying business logic.

Leveraging its self-developed Nebula Telecom Large Model, ZTE has built a multi-agent collaboration system trained on massive, high-quality communication-specific corpora to precisely address complex network O&M



Cross-domain
Business quality optimization
Service complaint processing
User experience assurance
Fault monitoring
Cross-domain Intelligence: VMAX-S & VMAX-AIMind & CEM
Agent
AI LLM engine
AI factory
Nebula model database
AI toolchain
Single-domain
Fault monitoring
Performance optimization
Quality optimization
Energy effiency optimization
Network change monitoring
Agent
Single-domain Intelligece: RNIA & CNIA & CIIA & ZENIC ONE
Network element
Wireless
Core network
Transmission & IP
Access
Agent
Intelligent computing board

▲ *Fig. 1 A panoramic overview of the Nebula Telecom Large Model empowering autonomous networks.*

challenges. With strong decoupling capabilities, ZTE's large model solution can flexibly adapt to multiple models. At the model foundation level, it not only supports ZTE's Nebula Telecom Large Model but is also compatible with outstanding open-source models in the industry such as DeepSeek. The core advantage of this solution is its seamless engine switching, allowing users to switch model engines across scenarios without impacting performance or compatibility.

This open architecture not only avoids the risk of technical path lock-in but also forms a healthy ecosystem of "foundation model competition and selection." Operators can flexibly choose foundation models according to business needs, utilizing both the generalization strengths of general-purpose models and the precision advantages of vertical domain models, ultimately realizing the vision of "the best model for the best scenario."

Multi-Agent Collaboration: Goal-Oriented "Task Autonomy"

Intelligent agents are key to fully unlocking the capabilities of large models, solving complex problems in communication networks independently and quickly through collaboration. As network O&M transforms toward large-model-driven operations, the paradigm is shifting from "human + machine" to "machine + human". Agents interact via the language programming interface (LPI), moving beyond the traditional API-based approach of addressing fixed scenarios, thereby improving their generalization capabilities.

Today's agents mainly complete tasks through dialogue and limited tool calls within fixed task flows. To enhance task autonomy, they need to support more advanced long-horizon autonomous planning and broader domain tool calling. Therefore, ZTE is integrating Manus-like capabilities into foundation models while continuously developing domain tools for networks to better cope with complex tasks.

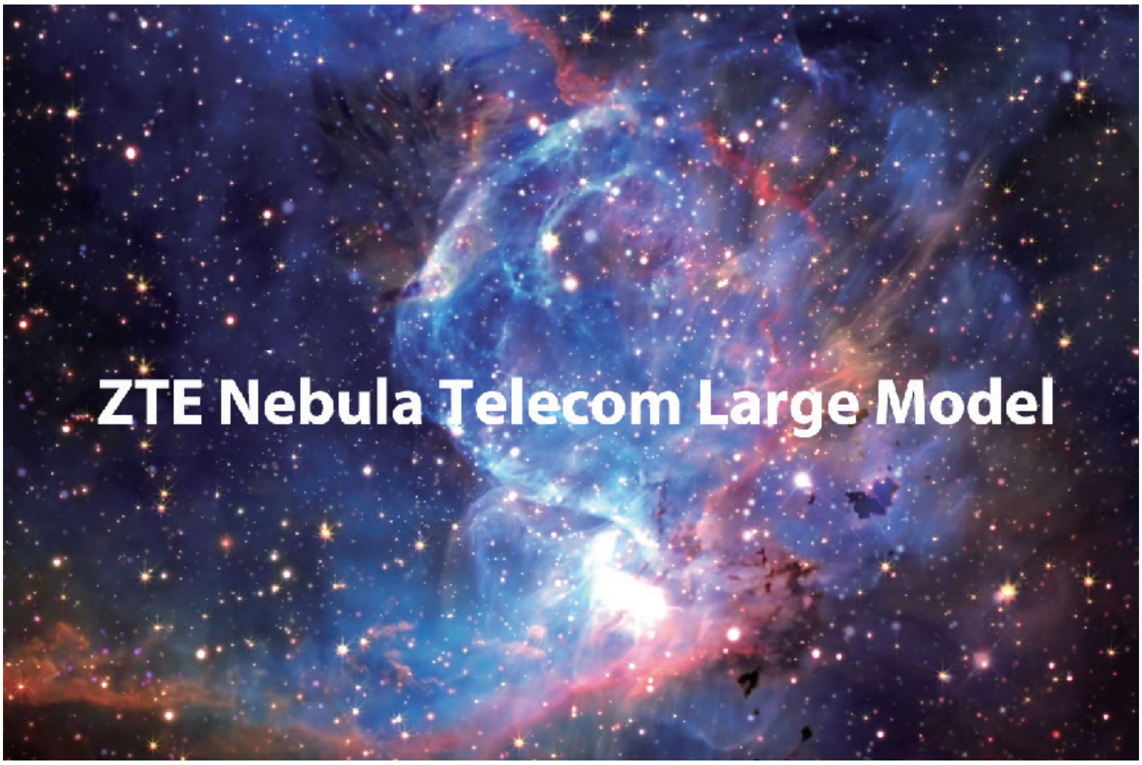
ZTE is actively building an agent-centric, iterative O&M system, where multiple atomic agents—both cross-domain and single-domain—are orchestrated based on scenarios to enable global agent collaboration. At the cross-domain layer, a multi-agent collaboration center orchestrates and flexibly invokes monitoring agents across and within domains. To ensure that agents can access the required capabilities when needed, atomic capabilities are developed in a composable manner. Specialized agents then chain these atomic capabilities according to scenarios, completing tasks through cross- and single-domain collaboration.

Knowledge Graph: Dynamic Integration for "Knowledge Dynamics"

With the growing scale and complexity of 5G networks, network fault monitoring and localization face severe challenges. Multi-source heterogeneous data (e.g. logs, alarms, topology, and device configurations) are scattered across isolated systems, leading to inconsistent semantics, significant format differences, delayed updates, and serious information silos. Traditional O&M methods, which rely on manual experience and static rule bases, cannot efficiently process massive unstructured data (e.g., work order records, maintenance logs) or to perform real-time root cause analysis in dynamic network environments. Addressing these challenges requires integrating multi-source heterogeneous data and adopting more effective representation methods.

Knowledge graphs provide a way to store and represent the relationships across such data and associated knowledge, consolidating scattered information—such as O&M experience, instruction manuals, and work orders—into a unified framework. They use graph search and enhanced large model reasoning to improve the efficiency and accuracy of fault localization. Both structured data and empirical corpus data in the communications field can be stored graphically, helping large models achieve more efficient utilization and accurate reasoning.

As the "memory hub" for telecom large models, knowledge graph faces two bottlenecks: data timeliness (due to dynamic changes in network configurations) and knowledge completeness (due



to variations across vendors' devices). Current knowledge graph technologies can partially structure knowledge but still heavily rely on manual annotation and predefined ontologies, This results in low automation, poor generalization, and lagging updates, failing to meet the real-time analysis and decision-making needs of fault propagation paths in complex network environments.

ZTE proposes an intelligent O&M framework that integrates multi-source heterogeneous data with LLMs, adopting a hybrid mode of "rule constraints + LLM reasoning" to build a dynamic knowledge graph. Unlike traditional methods relying on structured data, this approach leverages LLMs to extract implicit relationships from unstructured logs. It realizes real-time knowledge graph updates and dynamic generation of fault propagation trees through a stream processing engine. This provides both theoretical support and technical guarantee for accurate fault localization and root cause analysis in wireless networks, offering significant value in improving network reliability and O&M efficiency.

In-Depth Reasoning: Knowledge Enhancement for "Rigorous Inference"

The complexity of communication networks lies in long causal chains (from user perception to wireless access, transmission, and the core network) and multi-objective constraints (e.g., delay, energy consumption, and cost). General-purpose LLMs are prone to hallucinations and require the injection of domain-specific logical constraints through knowledge graphs. To improve inference accuracy, segmentation and reflection can be applied during large-model reasoning.

Phased reasoning consists of three sub-stages: hypothesis generation, knowledge verification, and iterative correction. In the first stage, the large model uses multimodal observation data—such as alarm logs, performance indicators, and topology status—together with pre-trained domain knowledge and pattern-recognition capabilities to generate an initial set of hypotheses. In the second stage, the hypothesis queue is input into the dynamic knowledge graph system for domain logical

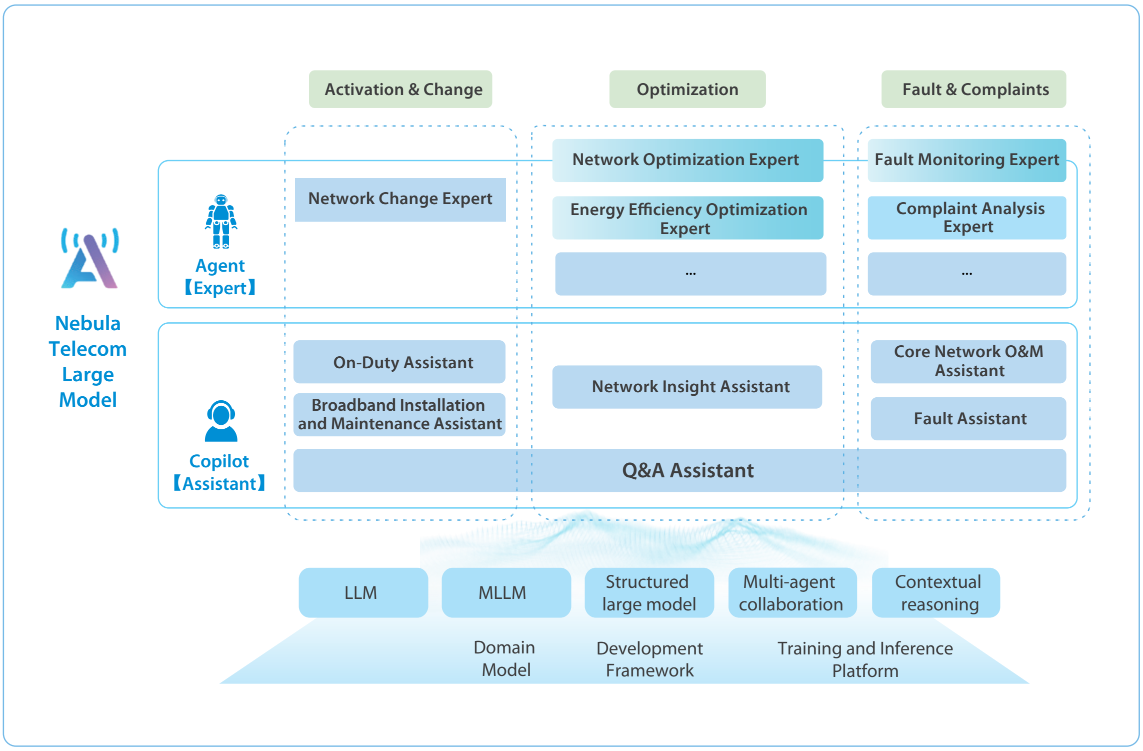

Activation & Change
Optimization
Fault & Complaints
Nebula Telecom Large Model
Agent【Expert】
Network Change Expert
Network Optimization Expert
Energy Efficiency Optimization Expert
...
Fault Monitoring Expert
Complaint Analysis Expert
...
Copilot【Assistant】
On-Duty Assistant
Broadband Installation and Maintenance Assistant
Network Insight Assistant
Core Network O&M Assistant
Fault Assistant
Q&A Assistant
LLM
MLLM
Structured large model
Multi-agent collaboration
Contextual reasoning
Domain Model
Development Framework
Training and Inference Platform

Fig. 2 Planning of the autonomous network copilots and agents.

constraint verification. In the final stage, a feedback reinforcement mechanism is established based on the verification results.

With an innovative model-adaptive, difficulty-graded distillation technology, ZTE's Nebula Telecom Large Model generates chain-of-thought (CoT) corpora that are manually proofread to enable the cold start of the foundation model. This stage endows the model with complete reasoning output capabilities. High-quality question-validator corpora are then used, along with reasoning-oriented reinforcement learning algorithms, to further enhance its performance in specific complex domains.

Harnessing the Large Model for Autonomous and Accurate Problem-Solving in Complex Scenarios

To advance toward high-level autonomy, the challenges of handling long-sequence, multi-dimensional, and structured data have been a central focus of technical research. Through refining the four basic capabilities, we enable copilots and agents for various scenarios to tackle these challenges (see Fig. 2).

Automated Closed Loop for Long-Process Scenarios

The core of an autonomous network lies in integrating technologies such as large models, intelligent agents, and digital twins, replacing manual judgment and operations with system-driven decision-making and execution to realize automated closed loops for end-to-end scenarios. Higher levels of autonomy place higher demands on these loops—shifting from a "best-effort" mode with fixed rules and manual fallback to addressing all long-tail scenarios and achieving comprehensive automated closed loops. This requires the O&M system to have generalization abilities, as well as enhanced autonomous process planning and problem-solving capabilities.

Taking cross-domain fault handling scenarios as an example, the complete process includes four stages: alarm detection, fault demarcation and localization, scheme execution, and effect verification. The traditional method requires manual data transfer and coordination across multiple systems, creating the risk of response delays. Intelligent transformation should eliminate these breakpoints and blockages, utilizing multi-agent collaboration to remove the need for manual operations. Through the flexible switching of foundation models and continuous iterations of multi-agent collaboration, the Nebula Telecom Large Model demonstrates Manus-like long-range planning capabilities and can enhance deep reasoning capabilities as industry models evolve.

- **Intelligent Multi-Dimensional Decision-Making**

Multi-dimensional decision-making is a key intelligent capability of communication networks. It requires balancing multiple target parameters such as bandwidth, delay, energy consumption, security, and cost in dynamic environments, and achieving globally optimal solutions through cross-layer collaboration and real-time computing. Traditional rule-based methods or single-objective optimization can no longer handle the nonlinear coupling problems in complex scenarios such as 5G/6G network slicing and edge computing. The introduction of model-based agents are reshaping the paradigm of network optimization by building a closed-loop cognitive system of "perception–inference–decision–verification."

Taking network optimization as an example, it is necessary to comprehensively consider multi-dimensional data (user distribution, traffic characteristics, and device status), establish a multi-objective optimization model, and generate optimal schemes under constraints such as delay, energy consumption, and cost. ZTE’s Nebula Telecom Large Model integrates knowledge graphs, structured data, and in-depth reasoning to obtain optimal solutions.

- **In-Depth Value Mining of Structured Data**

With massive network operation data, it is necessary to break through the limitations of traditional threshold-based alarms and identify hidden fault propagation chains by constructing alarm correlation models. At the same time, historical work order data can be mined to extract high-frequency solutions and predict potential network risks.

With the support of generative AI, vertical-domain corpora are evolving from auxiliary training sets into core knowledge carriers. Their role now goes beyond traditional feature engineering, serving as a “genetic blueprint” for domain cognition. ZTE’s Nebula Telecom Large Model overcomes challenges such as multi-source heterogeneous data fusion, professional knowledge distillation, lightweight reasoning, and graph search, ultimately enabling deep internalization and rapid retrieval of telecom knowledge through a systematic methodology.

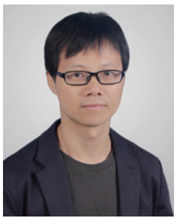
Summary and Future Prospects

The problem-solving, reasoning, and tool-invocation capabilities of telecom large mode-based agents are constantly improving. As accuracy meets production-level needs, the vision of "machine decision-making replacing manual decision-making" is gradually becoming a reality. This progress is grounded in the large model's rationality, complete logical reasoning, and ability to autonomously bridge process blockages and breakpoints. Therefore, the core task of domesticating large models lies in enhancing chain-of-thought reasoning within the communications domain, enabling precise multi-agent collaboration and tool use, and continuously expanding problem-solving capacity in real-world production scenarios. ZTE's Nebula Telecom Large Model is enhancing its capabilities in seamless engine switching, dynamic knowledge integration, and in-depth reasoning, extending its reach to complex scenarios and driving the network toward high levels of autonomy.

Looking ahead, with the commercialization of 5G-A and the arrival of 6G, future 6G networks will become AI-native, where AI applications will drive breakthroughs in intelligent perception, modeling, O&M, and resource scheduling. At the same time, network architectures, such as the data and control planes, will evolve to accelerate comprehensive AI deployment, realizing the mutual advancement of AI and 6G. ZTE TECHNOLOGIES

Intelligent Agent Factory:

Agile Way to Industrial-Grade Agents



Wu Jiangtao

LLM Planning Manager of Wireless and Computing Products, ZTE

Fu Linzhou

Senior Application Development Engineer for Wireless Networks, ZTE

When Manus gained global attention as the first universal agent, agent technology was undergoing a critical transition from laboratory research to industrial-scale production. To tackle three major pain points of traditional agents—long development cycles, inconsistent quality standards, and low asset reusability—the Intelligent Agent Factory was introduced. ZTE has developed the industry's first industrial-grade AI Agent platform that implements a "production-evaluation-optimization-closed-loop" workflow, driving agent development into the industrial era through standardized, modular production.

Core Architecture and Production Model of Agent Factory

The industrial-grade AI Agent Factory is based on a cloud-native architecture with underlying support for heterogeneous computing such as multi-vendor GPUs and inference cards, providing end-to-end support for the design, development, and operation of large-model-based agent applications (see Fig. 1).

RAG Workbench: Knowledge Assembly Workshop

The RAG workbench helps users quickly build applications powered by retrieval-augmented generation (RAG) technology. It enables effortless construction of knowledge bases through key steps such as corpus upload, intelligent segmentation, and embedding model selection, while also supporting evaluation and optimization with test datasets. Its knowledge retrieval feature provides customizable workflows, allowing flexible adjustments to retrieval strategies to enhance the response quality of RAG systems. With the RAG workbench, users can achieve end-to-end knowledge management and performance evaluation, rapidly developing high-accuracy, high-performance RAG applications.

WorkBridge Workbench: The Intelligent Hub Connecting Natural Language with APIs

WorkBridge is a development tool that bridges atomic capabilities and extends the capabilities of large models. It accurately maps natural language to existing APIs via natural language programming interface (LPI), enabling intelligent invocation—using language as command—to drive API execution. Its core features include:

- **LPI management:** Supports the creation, optimization, evaluation, and release of LPIs, ensuring accurate and reliable mapping between natural language and APIs.
- **Skill management:** Combines multiple LPIs into reusable business skills, providing full lifecycle management for these skills.

WorkBridge effectively lowers the technical barriers to development, accelerates intelligent application development, and sets a benchmark for engineering practices in natural language to API (NL2API) technology.

Agent Workbench: The Ultimate Assembly Station for Intelligent Agents

The agent workbench offers three approaches to meet diverse user needs for agent assembly: rapid development based on pre-built templates, AI-assisted intelligent construction, and customized professional development.

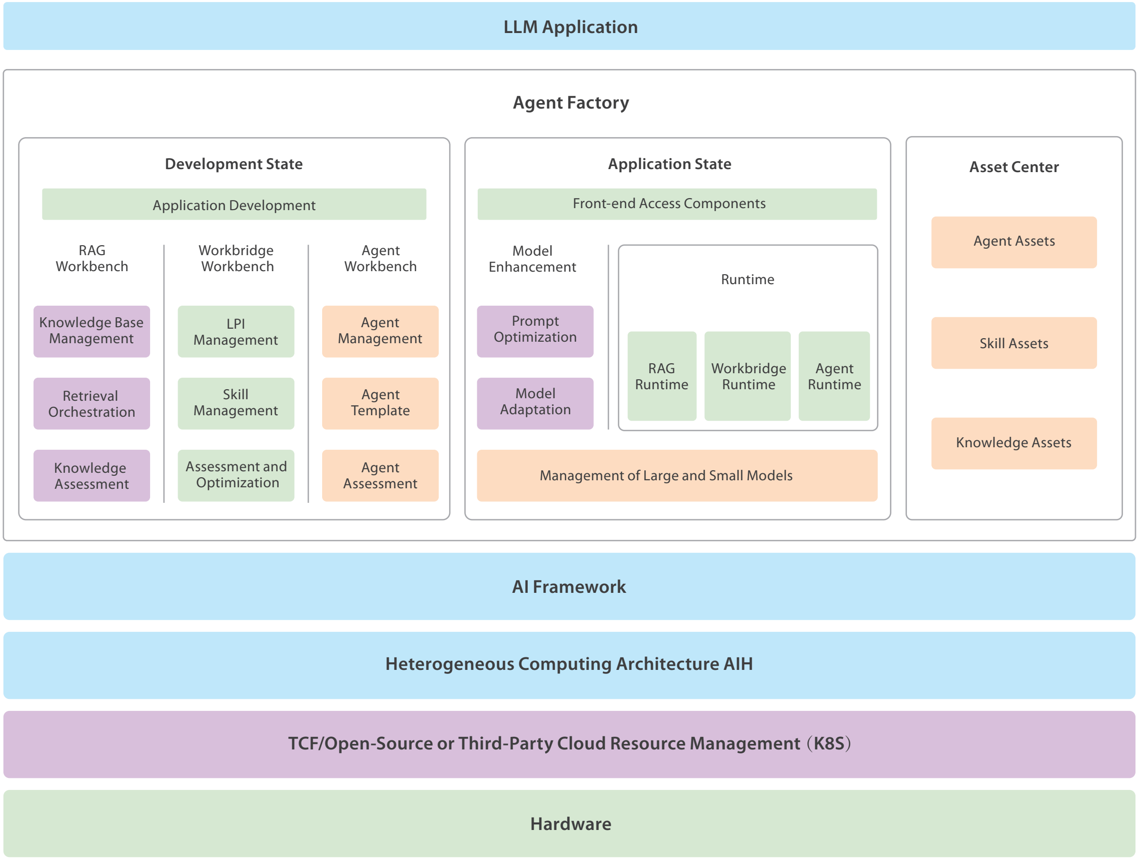


Fig. 1 Core technical architecture of the Agent Factory.

During the development process, the workbench provides a visual configuration interface and intelligent conversational assistance, enabling developers to quickly complete core configurations, such as agent role definition, knowledge base association, and tool integration. Developers can validate agent performance through manual testing or automatically generated test sets.

In addition, the workbench provides powerful evaluation and optimization capabilities, offering multi-dimensional metric analysis and comprehensive assessment reports to support continuous refinement. It also leverages AI to generate intelligent optimization suggestions, helping developers rapidly enhance agent performance.

Dual-State Collaboration: An Industrial-Grade Agent Production Pipeline

The dual-state model empowers industrialized agent production: in the "Development State," knowledge production, skill development, and agent assembly are collaboratively completed through three major workbenches (RAG, WorkBridge, and Agent), with the agents entering the asset center after rigorous testing. In the "Application State," the asset center enables one-click deployment, forming a standardized pipeline from raw materials to finished products. This pipeline balances quality and efficiency, providing reliable support for large-scale AI deployment.

In-Depth Practice: Crafting a Network Fault Monitoring Agent

Traditional monitoring systems suffer from delayed fault detection, low accuracy in root cause identification, and inefficient cross-system coordination. These challenges urgently require the application of AI technology to build an intelligent fault monitoring system, enabling a qualitative shift from passive response to proactive prevention.

ZTE adopts a progressive knowledge–skill–agent architecture to create a network fault monitoring expert.

Knowledge Engineering: Building the Intelligent Foundation

In the scenario of network fault monitoring, enabling the system to perform efficient and accurate interactive fault diagnosis requires the establishment of a knowledge system that covers fault definitions across multiple domains and scenario-based handling guidelines. The implementation involves two key aspects:

- **Knowledge base construction:** Supports importing from text, Word, and PDF files. The focus is on building a fault knowledge base (including fault definition standards, fault handling guidelines, alarm response manuals, etc.) and an equipment information database (mapping relationships between full and abbreviated names of equipment). Document import and intelligent segmentation are completed via the RAG workbench.
- **Retrieval mechanism design:** Adopts a hybrid vector + keyword retrieval model, utilizing the RAG workbench's visual workflow orchestration capabilities to intelligently re-rank retrieval results.

Capability Assembly: Unblocking the System's Meridians

Monitoring center engineers often need to frequently switch between multiple systems to handle faults. To address this issue, ZTE has encapsulated the APIs of the capability open platform (for retrieving alarms, performance metrics, logs, etc.), the work order platform, and the single-domain workbench (for topology queries) with natural language conversion. Using WorkBridge, APIs are converted into LPIs, and multiple LPIs are combined to form complete skills. Currently, 15 external system APIs have been encapsulated, supporting end-to-end closed-loop fault monitoring and handling.

Agent Production: Building a Digital Expert Team

ZTE has established a specialized "task force" for fault monitoring scenarios, comprising a fault identification agent, a fault analysis agent, a fault dispatch agent, and a report generation agent.

Taking the fault analysis agent as an example, its core function is to achieve fault demarcation and localization through multidimensional data analysis. In actual network operation and maintenance, fault analysis approaches vary across domains. For instance, addressing data network faults requires integrating aspects such as equipment in the computer room and transmission links, while network cloud faults necessitate layer-by-layer analysis from hardware to virtual layers and network elements. We design the demarcation and localization methods as a chain of thought stored in the knowledge base, then configure the agent with this knowledge base and equip it with skills such as alarm analysis and log analysis. After assembly, the agent undergoes iterative testing and optimization. Once it meets the standards, it is released to the asset center and deployed in the production environment.

Implementation Results

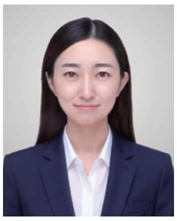
At a provincial operation and maintenance center, ZTE's fault monitoring expert "digital employee" operates 24/7, capable of accurately identifying faults within one minute and achieving a 91% accuracy rate in complex fault demarcation. Efforts are underway to vertically expand professional domains and horizontally broaden application areas, continuously enhancing the coverage of intelligent operation and maintenance.

In the future, the AI Agent Factory will evolve along two key directions: automation and specialization. For automation, AI technology will be leveraged to achieve an end-to-end closed-loop process from demand analysis to performance optimization, lowering the development threshold. For specialization, predefined agent template libraries will be constructed based on the communications field, enabling out-of-the-box deployment and accelerating the rollout of intelligent agent applications. As intelligent technologies advance, the AI Agent Factory is set to become the foundation for enterprise intelligent transformation, unlocking the inclusive value of AI. ZTE TECHNOLOGIES

Intelligent Agents and Multi-Agent Collaboration: Advancing Networks to New Heights of L4 Autonomy

Zhang Wenshuan

AI and Large Model Chief Planning Engineer, SDI Products, ZTE



Feng Yuan

IDA Project Large Model R&D Manager, ZTE

The "AI+" initiative has been included in China’s government work report for two consecutive years, with the 2025 report specifically advocating for the widespread application of large models. To drive AI technology towards commercial success, it is crucial to implement technology in practical scenarios and iterate cognitive paradigms. AI agents serve as the bridge between AI technology and its commercial application. In 2025, AI agents and multi-agent collaboration technologies are rapidly evolving, driven by both research and applications.

In network operation and maintenance, AI agents can unify the scheduling of large and small models alongside various network management system capabilities, creating a new paradigm for network O&M. TM Forum's "Autonomous Network: L4 Industry Blueprint—High-Value Scenarios", released in November 2024, introduces full-stack AI, updates the system architecture, and incorporates AI agents into the resource operation, business operation and commercial operation layers. The adoption of single-agent tools, together with cross-layer and cross-domain collaboration among multiple AI agents, will be key to driving the incremental evolution of autonomous networks towards L4 autonomy by 2030.

Empowering L4 High-Value Scenarios with Agents and Multi-Agent Collaboration

Building on TM Forum's high-value scenarios, ZTE has expanded its scope by identifying approximately 30 high-value scenarios for L4 advanced autonomous networks. These include handling personal service complaints, optimizing the quality of 5G private networks and IoT services, monitoring and handling network failures, and improving network performance.

To address these high-value scenarios, ZTE is focusing on developing six types of assistants and seven types of expert agents based on the Nebula Telecom Large Model and Nebula Telecom Specialized Models (such as the Signaling Large Model and Spatio-Temporal Large Model) (see Fig. 1).

Leveraging intelligent agents and multi-agent collaboration, ZTE has established several high-level autonomous network practices, including intelligent handling of network cloud failures, mobile network service complaints, and cross-domain failures.

Network Cloud Fault Intelligent Handling

Built on ZTE's cloud infrastructure intelligent analysis system (CIIA), the network cloud fault handling application utilizes multi-agent technology to dynamically decompose over 70 task with an accuracy rate of over 90%. This approach effectively handles complex domain tasks and and has been verified in the field. In practical applications, a telecom operator deployed the CIIA product, leveraging the Log Large Model's fault monitoring and diagnosis capabilities to enhance automation in on-site fault handling, reduce reliance on experts, and significantly save on workload. The process of identifying and diagnosing switch faults was reduced from over 140 minutes to less than 20 minutes.

Intelligent Handling of Mobile Network Service Complaints

ZTE's VMAX mobile service complaint solution

Fig. 1 The intelligent agent applications in ZTE's autonomous networks.

Copilot Assistants (6 categories, 20 products)

Product Name	User Role
Fault assistant	Monitoring engineer
Household broadband installation and maintenance assistant	Installation and maintenance engineer
Monitoring assistant	O&M engineer
Network insight assistant	O&M engineer
Q&A assistant	O&M engineer
Core network O&M assistant	O&M engineer

Agent Experts (7 categories, 21 products)

Product Name	Application Scenario
Network change expert	Network change
Network optimization expert	Network optimization
Fault monitoring expert	Fault monitoring
Complaint analysis expert	Complaint processing
Energy efficiency optimization expert	Energy efficiency optimization
Assurance expert	Key event assurance
Network insight expert	Network planning

introduces a complaint analysis agent that helps operators improve the quality and efficiency of complaint handling across four key stages: complaint reception, preliminary processing, complaint resolution, and quality control archiving. Compared to traditional methods, automated analysis significantly reduces the number of complaint tickets and shortens response times. As a result, the overall handling time for mobile service complaints is reduced by 50%, the interception rate for network issue complaints increases by 20%, fault localization accuracy reaches 84%, and first-line maintenance dispatching time is cut by 10%.

Intelligent Cross-domain Fault Handling

ZTE's fault monitoring expert, through the coordinated operation of a cross-domain analysis agent, a single-domain analysis agent, and a solution generation agent, seamlessly integrates fault identification, demarcation, localization, and scheduling execution. It also enables high-level collaborative analysis between the business layer and the network layer at both cross-domain and single-domain levels. In practical applications, a provincial mobile operator improved the IP network fault monitoring process with significant results: fault identification time was reduced from 5 minutes to real-time level of 1 minute, average repair time (MTTR) decreased by 8%, and the accuracy of fault demarcation by the large model significantly increased by 20%, reaching 91%.

Advancing Continuously to Tackle Challenges in Advanced Autonomy

L4 autonomy requires multi-agent collaboration across layers and domains to achieve end-to-end closed loops in complex scenarios, ultimately enabling distributed decision-making along with high-level flexibility and high-level scalability. Despite continuous advancements, challenges remain in multi-agent collaboration technology, the effectiveness of embedding applications into production, and the development of an ecosystem with protocols and platforms.

From a technological perspective, the multi-agent collaboration framework is still under development. Issues persist, such as the amplification of hallucinations among agents and the insufficient validation of decentralized communication mechanisms in applications. ZTE's Nebula AI Agent Engine supports cooperative, competitive, and mixed collaboration types, as well as rule-driven, role-driven, and model-driven collaboration strategies to pragmatically address these challenges step by step.

In terms of application, ZTE achieved significant verification results in 2024 in high-value scenarios such as energy conservation, cross-domain fault analysis, and network changes. These were set as benchmarks and replicated at scale. In 2025, ZTE will enhance value-effectiveness verification for key high-value scenarios, such as wireless network optimization and fault management in autonomous networks, embedding them into production processes and collaborating with partners to facilitate replication.

From an ecological perspective, although protocols and platforms related to AI agents are constantly emerging, the lack of unified standard poses a challenge to interoperability. ZTE continues to embrace open-source ecosystems and contribute to open-source communities to accelerate the maturity of open-source standards. In April 2025, during authoritative GAIA benchmark tests, ZTE open-sourced the Co-Sight Super AI Agent and topped the open-source framework list with an average score of 72.72. Meanwhile, the Nebula Agent Engine supports open-source protocols such as the model context protocol (MCP).

Looking ahead, ZTE will join hands with industry partners to build a comprehensive ecosystem for large models and AI agents. By overcoming technical barriers and enriching agent-centric high-value scenario applications for end-to-end autonomous networks, ZTE aims to advance networks to new levels of autonomy.

Structured Signaling Large Model: Facilitating the Intelligent Revolution at the Protocol Level

Chen Yan
Planning Specialist, Data Intelligence and Service Product Team, ZTE

Chen Xiangning
SDI Product Development Manager, ZTE

5G-Advanced (5G-A) is accelerating the intelligent transformation of high-value industries such as cloud gaming, the industrial metaverse, telemedicine, and vehicle-to-everything (V2X), raising network quality requirements from "usable" to "extremely reliable." However, the complexity of signaling interactions has grown exponentially, with a single user generating tens of thousands interactions daily. Traditional cross-domain data record (XDR) analysis suffers from cell loss (>30%) and limited scenario coverage (60%), forcing O&M personnel to manually parse raw signaling—often taking over eight hours on average. Additionally, complex faults (e.g., timing conflicts and protocol compatibility issues) are difficult to localize, impacting complaint resolution and network optimization.

ZTE has introduced a structured signaling large model that overcomes the barriers to intelligent protocol understanding, shifting from "manual signaling decoding" to "model-based protocol cognition." Leveraging machine cognition and end-to-end reasoning, the solution redefines the signaling analysis paradigm, significantly enhancing operators' intelligent O&M capabilities. Compared with traditional methods, it improves signaling analysis efficiency by 80%, raises fault localization accuracy to over 95%, and enables operators to deeply participate in scenario-specific optimization during the fine-tuning phase—helping build an intelligent user experience assurance system for the 5G-A era.

Innovative Signaling Analysis Solution Based on Structured Large Model

The solution establishes an end-to-end intelligent signaling analysis system through a collaborative architecture comprising data, model, and service layers, enabling a leap from raw signaling parsing to intelligent reasoning analysis.

Three-Layer Architecture Design

The architecture is shown in Fig. 1:

- **Data layer:** Supports efficient encoding of raw signaling, covering over 100 types of network element interfaces and related protocols, including 5GC, EPC, and IMS, ensuring the integrity and consistency of signaling data.
- **Model layer:** Integrates structured signaling large models and large language models (LLMs), providing pre-trained foundational models, chain-of-thought fine-tuned models, and intelligent reasoning capabilities. This enables accurate understanding of complex signaling logic and enhances fault diagnosis and anomaly analysis.
- **Service layer:** Incorporates real-world business scenarios to provide signaling visualization, anomaly detection, automated analysis reports, and intelligent Q&A, empowering operators to achieve intelligent network O&M.

Structured Signaling Large Model: Evolving From Rule-Driven to Intelligent Cognition

The solution uses structured large models to learn original signaling interaction patterns and integrates a chain-of-thought mechanism based on expert experience, enabling the model to autonomously parse signaling.

Signaling is essentially the "language" between network devices, characterized by fixed formats and low information dispersion, making it more suitable for LLM-based methods than natural language.

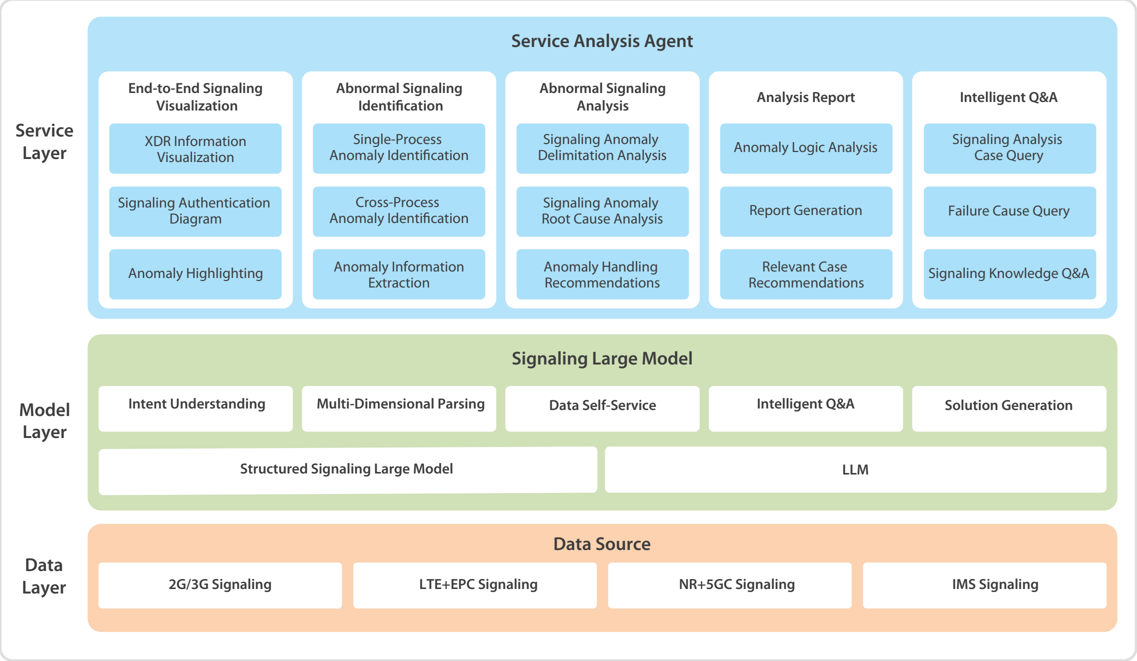

Service Layer
Service Analysis Agent
End-to-End Signaling Visualization
XDR Information Visualization
Signaling Authentication Diagram
Anomaly Highlighting
Abnormal Signaling Identification
Single-Process Anomaly Identification
Cross-Process Anomaly Identification
Anomaly Information Extraction
Abnormal Signaling Analysis
Signaling Anomaly Delimitation Analysis
Signaling Anomaly Root Cause Analysis
Anomaly Handling Recommendations
Analysis Report
Anomaly Logic Analysis
Report Generation
Relevant Case Recommendations
Intelligent Q&A
Signaling Analysis Case Query
Failure Cause Query
Signaling Knowledge Q&A
Model Layer
Signaling Large Model
Intent Understanding
Multi-Dimensional Parsing
Data Self-Service
Intelligent Q&A
Solution Generation
Structured Signaling Large Model
LLM
Data Layer
Data Source
2G/3G Signaling
LTE+EPC Signaling
NR+5GC Signaling
IMS Signaling

Fig. 1 Three-tier architecture of the structured signaling large model.

As a result, by deeply learning network protocol features, the model achieves end-to-end signaling parsing, protocol conflict detection, and abnormal pattern recognition.

Three Core Models Enabling End-to-End Intelligent Signaling Analysis

The solution establishes a comprehensive intelligent signaling analysis architecture through the collaborative optimization of encoding models, projection models, and domain decoding models. It also provides operators with opportunities for deep participation during the fine-tuning phase to ensure precise adaptation of the models to real-world network environments.

- **Signaling encoding model:** Built on an improved Transformer and a multi-level attention mechanism, it analyzes the byte-level content and hierarchical structures of signaling protocols, providing accurate semantic representations of signaling.
- **Projection model:** Constructs a mapping between the signaling semantic space and the business rule space, unifies data in different protocol formats, achieves automatic feature alignment, and reduces adaptation complexity.
- **Domain decoding model:** Enhances signaling reasoning capabilities and supports various intelligent O&M applications by embedding 3GPP protocol standards and integrating signaling processes, messages, and business scenarios.

Applying RAG for Precise Signaling Knowledge Reasoning

By integrating retrieval-augmented generation (RAG), the solution significantly improves the accuracy and efficiency of signaling parsing:

- **Retrieval optimization:** Introduces pre-retrieval routing, query rewriting, index optimization, and re-ranking techniques to greatly enhance the efficiency and relevance of signaling data retrieval.
- **Reducing hallucinations in large models:** Combines foundation models with external knowledge sources to optimize the information generation process, ensuring the accuracy and interpretability of reasoning results.

Application Empowerment: Complaint Analysis Agent

As 5G evolves toward 5G-A and future 6G, network services are becoming increasingly diverse, making user experience assurance a core demand. Faced with

> In 2025, ZTE completed key technology verification of an intelligent complaint analysis agent based on a structured signaling large model in Jiangsu, marking a milestone in intelligent complaint handling.

the complexity of new services and scenarios, traditional user complaint handling models struggle to provide rapid response and precise issue identification.

ZTE has developed a complaint analysis agent powered by the structured signaling large model, introducing a digital employee—the "user complaint handling expert"—to achieve intelligent and automated end-to-end signaling analysis, helping operators improve efficiency and reduce costs.

- **Intelligent complaint Q&A: Lower the threshold for signaling analysis and improving response efficiency**

Complaint-related signaling data is vast and complex, with intricate protocol flows. Manual analysis relies heavily on expert experience, and fragmented knowledge leads to low processing efficiency. By leveraging a structured signaling large model, ZTE has developed an intelligent Q&A engine that provides signaling analysts with immediate and accurate knowledge support through human-machine collaborative interaction and historical knowledge accumulation. This lowers the barrier to signaling processing and enhances analysis efficiency.

- **Automated complaint signaling analysis: Accurately pinpoint anomalies and swiftly provide solutions**

Users can click on the agent entry point or directly input their issue intent to trigger the automated complaint analysis process. The agent enables automatic identification of abnormal signaling and intelligent screening, parses and visualizes signaling processes for intuitive root cause analysis, provides standardized signaling explanations to lower technical barriers, and offers handling suggestions and failure cases to support precise decision-making.

- **Intelligent complaint analysis report generation: Free up O&M resources and improve service quality**

The solution integrates RAG technology and consolidates multi-domain and multi-type data to generate comprehensive and accurate complaint analysis reports. Report generation time is significantly reduced, minimizing manual effort and enhancing the standardization and automation of the complaint handling process.

In 2025, ZTE completed key technology verification of an intelligent complaint analysis agent based on a structured signaling large model in Jiangsu, marking a milestone in intelligent complaint handling. Moving forward, ZTE will strengthen its focus on 5G-A networks, and launch intelligent complaint resolution solutions for both ToC (individual users) and ToB (enterprise users), helping operators build a more efficient, accurate, and intelligent complaint management system.

As a leading global provider of telecommunications equipment and network solutions, ZTE will continue to drive innovation and advance digital and intelligent networks, delivering superior network experiences for customers worldwide. ZTE TECHNOLOGIES

Network Insight Agent: Enabling AI-Driven Multi-Dimensional Insights and Solution Generation

Zhao Xin

Wireless Network AIOps Product Manager, ZTE

Yin Jianhua

Chief Engineer of Wireless Network AIOps Pre-research, ZTE

Yan Haibo

Wreless Network Product Planning Manager, ZTE

As a core application of the large model for wireless AI for IT operations (AIOps), the Network Insight Intelligent Agent provides a traffic stimulation solution for wireless access networks. It delivers in-depth insights into network structure, coverage, capacity, device health, and other dimensions. Leveraging generative AI, the solution generates query strategies, insight summaries, solution recommendations, and multi-dimensional charts—showcasing the powerful capabilities of ZTE's Nebula Telecom Large Model. The agent supports users in efficiently understandig network insights and solution demands across different stages, application scenarios, and objectives through natural language interaction, enhancing the efficiency of network analysis and solution generation in network planning and optimization.

AI-Driven Applications for Intelligent and Efficient Network Data Analysis and Solution Generation

Traditional network analysis and solution output require operations and maintenance personnel to query large volumes of network data, process it, and provide network planning and optimization recommendations, relying heavily on manual analysis and expert experience. With the introduction of large language models (LLMs), key algorithms and technologies such as intent understanding, retrieval-augmented generation (RAG), NL2API, NL2SQL, NL2Code, and long short-term memory (LSTM) enhance the intelligence of data analysis and solution generation.

The Network Insight Agent brings the following user values:

- **Intent understanding:** Leverages the language comprehension and reasoning capabilities of large models to generate network insights through natural language dialogue.
- **Report generation:** Flexibly generates text summaries, chart presentations, and solution suggestions based on user demands or preferences, creating materials for communication and reporting.
- **User adaptation:** Adjusts output based on user feedback to better align with their demands.
- **Professional enhancement:** Through features like knowledge Q&A and guided questioning, users learn while using the system—clarifying concepts, improving problem description skills, and enhancing their problem-solving abilities.

The six key functional features of the Network Insight Agent are as follows:

- **Multi-agent collaboration:** Multi-dimensional network insight agents, including expert agents for structure, coverage, capacity, and health, collaborate in traffic stimulation scenarios. They provide multi-dimensional insights, offering conclusions and solution recommendations.
- **On-demand data query:** Provides eight-dimensional insights for wireless access networks, including structure, coverage, traffic, load, health, and and three additional dimensions (assets, performance, and energy efficiency) to be supported in the future. Users can query statistical data on demand through natural language interaction, enabling a network health check.

> **The Network Insight Agent not only addresses inefficiencies and information asymmetry in traditional operational models but also delivers breakthroughs in multi-dimensional data analysis and expert agent collaboration.**

- **Solution generation:** Generative AI, based on expert knowledge bases and insight data, generates insights and solutions.
- **Personal agent creation:** Users can create personalized combinations of intelligent agents based on their preferences, enabling scene orchestration to cover multiple scenarios.
- **Query strategy generation & chart creation:** Utilizing the semantic understanding, generation, and logical reasoning capabilities of large models, this feature automates the generation of query strategies, insight conclusions, and solution recommendations.
- **Input association:** When users input simple terms like "5G" or "coverage," the Network Insight Agent automatically associates related questions, such as "How many high-load 5G cells are there?" and "What is the distribution of weak coverage cells?" This simplifies user input while demonstrating the product's capabilities.

Multi-Agent Collaboration: In-Depth Application of the Network Insight Agent in Network Planning Scenarios

An intelligent agent provides strong operational capabilities at the core of large models, unlocking their full potential. With clear objectives, the agent can independently think and take actions to achieve these goals. It breaks down the task into detailed steps, utilizing feedback from external sources and its own reasoning to generate prompts to accomplish the goal.

For example, in a network planning application, if a user asks, "What is the network situation around the Datang Everbright City block?" the agent will decompose this task into four main steps: analyzing the sites in the Datang Everbright City area, determining if the sites are operating stably, checking if the coverage meets the requirements, and evaluating if the network capacity is sufficient.

The multi-dimensional network insight agent first gathers expert agents specializing in network structure, network coverage, network capacity, and device health to analyze the network situation in the area.

- **Step 1:** The Structure Expert Agent filters sites from the configuration table.
- **Step 2:** The Health Expert Agent calls relevant interfaces to check for service outages, alarms, and hidden network element issues.
- **Step 3 and Step 4:** The Coverage Expert Agent and Capacity Expert Agent select the tools for the coverage and capacity insights to execute necessary tasks.
- **Final step:** The multi-dimensional network insight agent consolidates the information from all expert agents and returns the final response to the user.

To support these user-interactive business

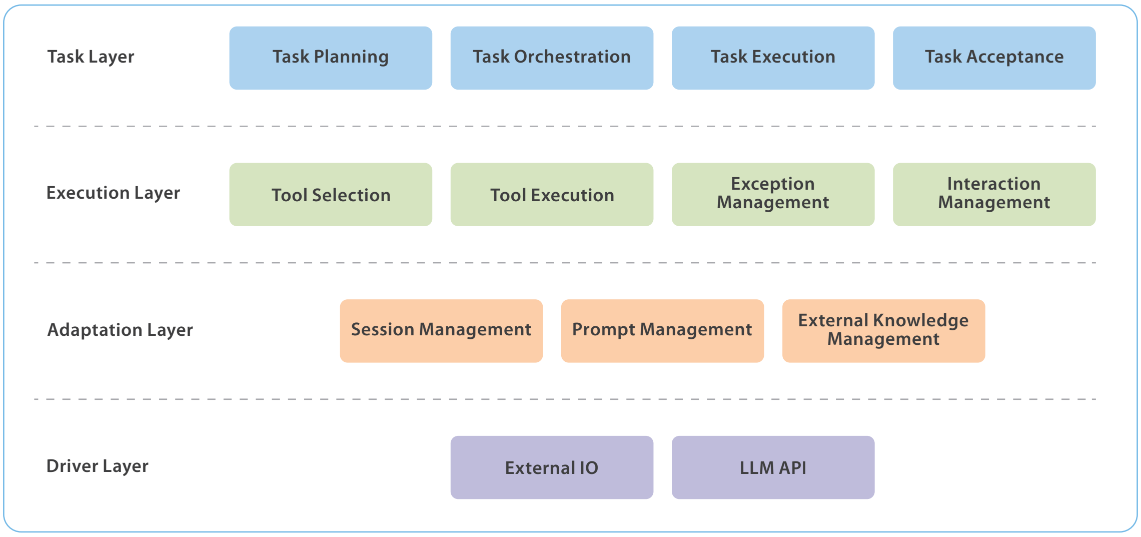


Fig. 1 Layered concept design of the Network Insight Agent.

processes, the Network Insight Agent is designed using the layered structure shown in Fig. 1.

- **Task layer:** The agent breaks down the given task and conducts overall planning and task orchestration.
- **Execution layer:** Based on the task designed, the agent identifies suitable tools and provides the necessary input information to execute the tool's actions.
- **Adaptation layer:** Includes session management, prompt management, and external knowledge management, while also providing both short-term and long-term memory capabilities.
- **Driver layer:** Includes management modules for accessing vector databases, API libraries, and LLM API modules for accessing large models.

The network planning scenarios utilize the multi-dimensional network insight agent and the collaboration of multiple single-dimensional expert agents. As the core agent, the multi-dimensional network insight agent, is capable of task planning, organizing expert analysis, summarizing expert recommendations, and proposing improvement measures. Multiple single-dimensional expert agents query network data in their respective domains to generate solutions.

Personifying the specialized roles of agents and utilizing APIs, knowledge bases, and online data from various dimensions can significantly reduce the hallucination issues of large models and improve the reliability of the generated solutions. Additionally, the collaboration of multiple agents enables problem-solving in multi-goal and complex scenarios.

Currently, the Network Insight Agent, along with ZTE's AIOps system, has been commercially deployed at 24 major sites (provincial-level network management systems) across three major telecom operators in China. During the MWC 2025, held in early March, the Network Insight Agent showcased ZTE's large-model applications on the wireless side and conducted a live interactive demonstration. In the second half of 2025, this application will be implemented in the wireless network management systems of overseas operators, including AIS in Thailand.

The launch of the Network Insight Agent and other agent series marks a new era of intelligence in ZTE's network O&M. Through deep learning and large model technology, it not only addresses inefficiencies and information asymmetry in traditional operation modes but also delivers breakthroughs in multi-dimensional data analysis and expert agent collaboration. Going forward, the Network Insight Agent will evolve to support more precise queries, finer-grained network analysis, and more diverse application scenarios. ZTE TECHNOLOGIES

Wireless O&M Agent: Shaping the Future of Network Assurance

Shen Yuan

Chief Planning Engineer of RAN Products, ZTE

Shen Yi

RNIA Project Manager, ZTE

The communications industry is accelerating its intelligence transformation to meet the challenges of growing network scales and complicated service scenarios. Traditional network O&M, which relies on human experience, suffers from low efficiency, slow responses, and complex operations, making it inadequate for assuring high-value areas such as large shopping malls and major activity venues. By integrating AI with large models, ZTE's Nebula Telecom Large Model and its derived O&M agent technology have created a new paradigm for network assurance, driving the industry's intelligent upgrade. For high-value area assurance, ZTE and its partners, in multiple commercial pilots in China, reduced manpower investment by 83% and increased efficiency by five times through its O&M intelligent agent solution, marking a major step toward intelligent network assurance.

Limitations of Traditional Assurance Models

Traditional network assurance models face four key challenges:

- **Low O&M efficiency:** Slow alignment between network resource configurations with service requirements, hindering quick response to bursty traffic or complicated scenarios.
- **High expertise requirements:** Assurance strategies rely on expert experience, requiring front-line personnel to master multiple systems and driving up training costs.
- **Policy and objective disconnect:** Network configurations focus more on troubleshooting, with weak connection with business goals such as user experience optimization.
- **Insufficient dynamic response:** Frequent policy adjustments are needed in important scenarios, but low automation makes real-time optimization difficult.

ZTE's Wireless O&M Agent Solution

ZTE's Unified Management Expert (UME) wireless O&M agent (Fig. 1) is an agent application developed by ZTE for wireless network assurance scenarios based on the large model technology. It supports diverse assurance scenarios, including concerts, sports events, emergency assurance, and tidal traffic during daily O&M operations. Built on the Nebula Telecom Large Model, it adopts a modular design with four core modules:

- **Perception module:** Collects real-time network, user, and external event data, providing decision inputs through feature extraction and data cleaning.
- **Large model module:** Acts as an intelligent brain, analyzing intents and generating assurance policies with natural language processing (NLP) and knowledge graph technologies. Assurance can be triggered via calendar or email, enabling automatic scenario identification.
- **Planning & execution module:** Invokes network atomic capabilities (such as resource scheduling and parameter adjustment) to perform closed-loop operations based on reinforcement learning optimization policies.
- **Feedback learning mechanism:** Improves policy accuracy and adaptive capability through continuous data training.

The UME O&M agent integrates generative AI, multi-agent collaboration, and retrieval-augmented generation (RAG) to mitigate hallucinations in large models and ensure reliable decision-making. It offers the following core functions:

- **Full scenario coverage:** Supports major events

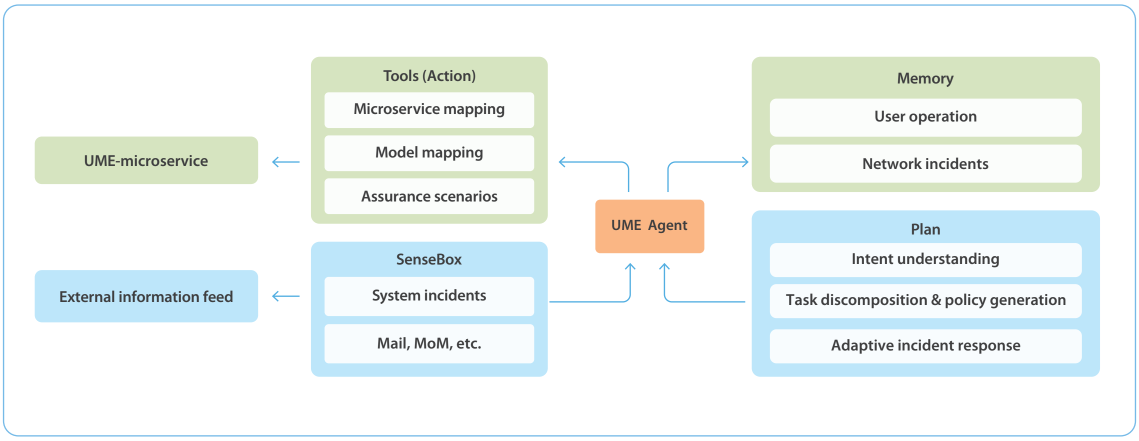


Fig. 1 Technical architecture of the UME O&M agent.

(e.g. concerts, sports), provides emergency assurance, and manages daily tidal traffic.

- **Intelligent generative interaction:** Triggers workflows through natural language instructions (e.g., "provide network assurance for the XX shopping mall during the evening rush hour"), automatically generates and executes policies, boosting operational efficiency fivefold.
- **Dynamic optimization and self-adaptation:** Monitors network indicators (e.g., user count, PRB usage, and interference levels) in real time and dynamically adjusts policies to deal with traffic fluctuations. In the Hangzhou Olympic Center concert project, network traffic prediction accuracy improved by 20%, greatly enhancing assurance efficiency.

ZTE's O&M agent solution is not only a technological breakthrough but also a reconstruction of O&M models.

- **Technology integration and innovation:** Deeply integrating large models with telecom knowledge to address traditional AI’s limitations in structured data processing, such as improving fault diagnosis accuracy with the RAG signaling knowledge base.
- **Ecosystem synergy:** ZTE works with operators and industry partners to build agent technology standard and scenario libraries, accelerating large-scale deployment.
- **Business value extension:** Beyond cost reduction and efficiency gains, it enables experience-driven operations and differentiated services (e.g., dedicated assurance for live streaming and cloud games) for operators, opening up new revenue opportunities.

Application Pilot and Results

ZTE selected Wushang Dream Times Square, the world's largest shopping mall in Wuhan, as the verification site, covering 31 physical cells and 60 logical cells. The mall averaged over 1,800 daily users, with holiday traffic reaching three times the usual volume. The traditional solution required six man-days, while the agent reduced resource query time from two minutes to 15 seconds and fault location time from 15 minutes to one minute.

- **Stable network indicators:** User-perceived rate rose 15%, interference dropped 30%, and PRB usage stayed healthy.
- **Efficiency breakthrough:** Tasks are delivered through natural language instructions, with policies automatically generated and executed, reducing manpower costs by 83%.
- **Economic benefits:** Saved O&M costs can be reinvested into the network, helping operators explore new business models such as experience-driven operation.

With the evolution of 6G and general computing integration, the O&M agent will be further upgraded to “all-domain Autonomous Networks”, empowering everything from network assurance to business innovation and setting a benchmark for the industry’s intelligent transformation.

Core Network O&M Agent Boosts Efficiency for L4 Autonomy

He Wei

Chief Engineer of Intelligent O&M Product Planning, ZTE

With the rapid deployment of 5G networks and increasingly diverse service scenarios, the operational complexity of the core network has grown sharply. Traditional O&M models, which rely on manual expertise, struggle to meet demands such as dynamic adjustments of large-scale networks, fault prediction, and self-healing. In this context, autonomous networks powered by large language models (LLMs) and intelligent agents have emerged as a key solution to enhance intelligent O&M for 5G core networks.

The Use of Intelligent Agents in O&M

Agent technology has been widely applied in telecom intelligent O&M. In the core network domain, typical application scenarios include fault management and complaint-handling agents.

- **Fault management agent:** The agent perceives real-time alarm information or processes fault tickets dispatched by external pipelines, detecting fault anomalies, analyzing root causes, providing resolution feedback, and offering handling suggestions. Fault tickets can be automatically responded to. Additionally, leveraging models trained on historical fault-handling experiences and combining them with real-time data monitoring and analysis, the agent can predict potential faults, issue early warnings, and help reduce the probability of fault occurrences.
- **Complaint-handling agent:** As user demands for superior experience continue to grow, complaints in networks have become increasingly diverse, involving potential end-to-end issues from terminals to wireless, bearer, and core networks. The lengthy resolution process remains a major challenge. The complaint-handling agent can analyze complaint tickets, including user subscription information, network configuration details, and signaling interaction data. Leveraging large model technology, the system identifies problems, pinpoints root causes, provides suggestions, and automatically closes tickets, enabling fully automated complaint handling, significantly reducing fault resolution time, and improving efficiency.

Architecture and Key Technologies of ZTE's Core Network O&M Agent

System Architecture

The system architecture of ZTE's core network O&M agent encompasses five layers: network, data, model, application, and digital twin (see Fig. 1).

- **Network layer:** Serves as the foundation of network operations and includes the existing atomic network elements within the core network.
- **Data layer:** Provides high-quality corpus data for decision-making in the upper layers.
- **Model layer:** Uses large and small models as intelligent engines to create a composable, orchestrated, and self-iterating intelligence foundation.
- **Application layer:** Orchestrates various application capabilities to meet O&M needs of diverse scenarios.
- **Digital twin:** Constructs a "business model + twin application" architecture to facilitate business innovation and network optimization.

Key Technologies

To realize these capabilities, the core network O&M agents employ several key technologies:

- **Graph-RAG:** Graph-based retrieval-augmented generation (Graph-RAG) integrates knowledge graphs with RAG to enhance LLM reasoning, overcoming traditional RAG's limitations in

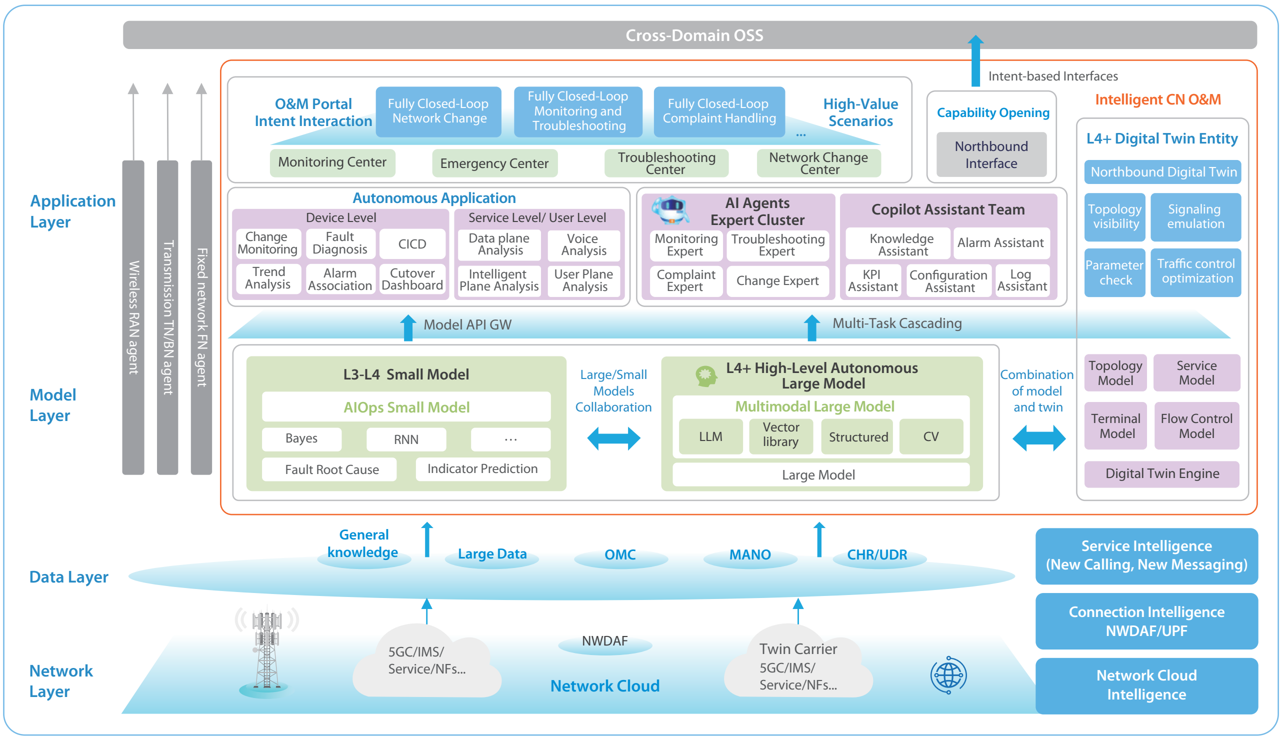


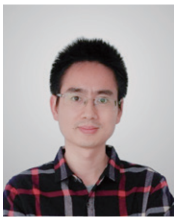
Fig. 1 ZTE's core network O&M agent system architecture.

complex queries and multi-hop reasoning.

- **AI Agent:** Multi-agent collaboration enables multiple agents to communicate and cooperate in a shared environment to accomplish common goals. Each agent has a certain degree of autonomy and intelligence to perceive, decide, and act based on environmental information. This collaboration allows the entire system to benefit from their complementary strengths, resulting in more efficient and intelligent decision-making. Based on this architecture, specialized agents—such as knowledge experts, fault experts, on-duty experts, and complaint experts—can be created to jointly build an intelligent O&M system.
- **MCP:** Model context protocol (MCP) standardizes how LLMs or generative AI systems understand, store, and utilize contextual information. It establishes a unified communication interface between AI models and external data sources or tools, covering capabilities such as context window management, multi-turn dialogue maintenance, and dynamic context updates. MCP functions like a USB interface of AI: any compliant AI model or tool can achieve fast "plug-and-play" integration, without the need for separate interface programming or programming language constraints. Compared with early function calling in large models, it significantly improves interaction modes, capability definition, protocol standardization, and ecosystem openness.
- **Digital Twin:** Digital twin technology uses discrete event simulation algorithms and related techniques to construct a digital twin model of the core network. This model supports O&M through visualization, simulation, predictive analysis, and strategy feedback, enabling qualitative and quantitative analysis at low cost. It facilitates the transition of decision-making from human-led processes to machine-assisted and even machine-autonomous operations, accelerating the evolution toward advanced autonomous intelligent O&M, and ultimately enabling full closed-loop automation.

The development of intelligent agents is shifting O&M models from "automation" to "autonomy," though key challenges in reliability, security, and collaboration remain. With breakthroughs in intelligent agents and digital twins, intelligent O&M is expected to achieve widespread implementation of L4 autonomy within the next five years, laying the foundation for highly autonomous intelligence in the 6G era. Continuous innovation, scenario-driven practice, and sustained R&D, are essential to accelerate this process.

Closed-Loop Network Change Agent: An O&M Innovation



Ou Xuegang

ITN Product Planning Manager, ZTE

In the digital era, networks are the backbone of enterprise operation and social development. As services expand and technologies evolve, network changes have become crucial for efficient network operation and service assurance. Traditional network change O&M models can't cope with complicated network environments and service demands. The rise of AI large models has introduced new solutions, enabling a closed-loop network change agent poised to transform traditional O&M.

Traditional Network Change O&M Pain Points

The traditional network change O&M system has many pain points that significantly restrict the efficiency and stability of enterprise network management.

- **High threshold for O&M skills:** Traditional network change O&M system relies heavily on O&M personnel for solution design and review. This requires deep knowledge of network topology, device configuration and management, and a precise understanding of network demands in different service scenarios. The shortage of qualified O&M professionals increases recruitment pressure and drives up labor costs.
- **Lagging automation:** In traditional network change O&M, core processes, such as pre-change inspection and configuration generation, rely almost entirely on manual operations, which are inefficient and also prone to human error. Without automated tools, it is difficult to fully and accurately evaluate the change impacts or identify potential risks in advance, increasing uncertainty during network changes.
- **High risk of service interruption:** In traditional O&M, the testing and verification phase, involves numerous complex steps, and manual verification is highly inefficient. If an issue is found during the test and rollback is required, manual operations may exceed the preset operation window, causing large-scale service interruptions, economic losses, and reputational damage to the enterprise. Manual verification is also subjective and limited in scope, making it difficult to detect subtle yet potentially critical issues that may affect long-term network stability.

Closed-Loop Network Change Agent Solution

The closed-loop network change agent, powered by AI large models, generates network change solutions based on the agent architecture, makes trustworthy decisions based on digital twins, and completes the closed-loop process—pre-change inspection, change simulation, change execution, change verification, and post-change monitoring—through natural language concatenation.

- **Automatic/semi-automatic change solution generation:** With powerful data analysis and processing capabilities, the AI large model intelligently orchestrates network-related APIs to generate dedicated tools for pre-change inspection and post-change verification. These tools automatically detect network status, collect key data, and compare the data with preset standards to quickly identify potential problems, providing reliable early-stage evaluation and later-stage verification. The solution also supports natural language interaction, allowing O&M personnel to easily communicate with the agent and input network change demands and instructions. The agent converts these inputs into executable steps, seamlessly connects all steps, and automates the entire process, significantly reducing operational difficulty and

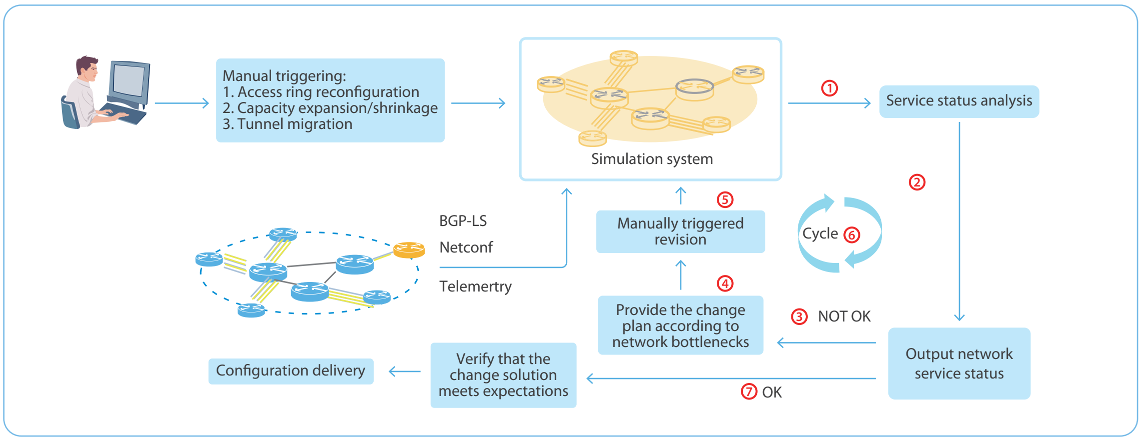

Manual triggering:
1. Access ring reconfiguration
2. Capacity expansion/shrinkage
3. Tunnel migration
Simulation system
①
Service status analysis
②
BGP-LS
Netconf
Telemertry
⑤
Manually triggered revision
Cycle ⑥
④
Provide the change plan according to network bottlenecks
③ NOT OK
Output network service status
Configuration delivery
Verify that the change solution meets expectations
⑦ OK

Fig. 1 Network change simulation.

communication costs for O&M personnel.

- **Trustworthy decision-making mechanism:** Leveraging digital twin technologies, the agent simulates network changes in a virtual environment (Fig. 1). It identifies potential issues such as network congestion and device compatibility conflicts, providing a scientific and reliable basis for making network change decisions and ensuring the security and reliability of the change operation from the outset.
- **Reliable implementation guarantee:** AI large models construct an execution chain of thought, integrating key steps such as pre-change checks, configuration changes, and post-change verification. Based on this chain, the agent can automatically and systematically perform operations according to preset standards and procedures, avoiding confusion and errors that may occur in manual operations.
- **Atomic operation mechanism:** The solution encapsulates the entire change execution process as an atomic operation. All operations are either completed in full or automatically rolled back to their pre-change state if a fault occurs, preventing network faults and data corruption caused by partial execution, and greatly improving execution stability and reliability.

Highlights of Network Change Agent

- **Improvement in fault prevention, control, and efficiency:** The closed-loop network change agent delivers breakthroughs in fault prevention, control and efficiency. Through automated and intelligent operations, it shortens traditional O&M change time from over 14.5 hours to just 4 hours, improving efficiency by more than 70%. This enables enterprises to complete network changes faster to meet rapid service development needs while minimizing service interruption windows, lowering risks, and improving both enterprise network service quality and user experience.
- **Trustworthy decision-making and steady execution:** Powered by a digital twin base, the agent ensures trustworthy decision-making. By simulating network changes in a virtual environment, the agent identifies obvious potential network faults and subtle performance impacts in advance, providing reference for decision-making. Robust change execution is guaranteed through an atomic operation mechanism that eliminates risks of major faults from operational errors or partial execution. In case of incidents, configurations can be quickly rolled back to restore the network to a stable state, reducing risks and losses from network changes.

The closed-loop network change agent represents a major shift in network O&M. It addresses long-standing challenges in traditional network change O&M through an intelligent and automated approach, greatly improving efficiency, reducing risks, and enhancing management quality. As AI technologies advance, the agent will expand into broader domains and scenarios, deeply integrate with IoT and cloud computing, and drive further innovations and breakthroughs in network O&M. ZTE TECHNOLOGIES

Discussion on Home Knowledge Base and Intelligent Agent

Wang Meng

Smart Home Product Planning Manager, ZTE

In 2025, AI has made significant strides, especially in user experience. It has moved beyond research and is now accessible to end users. The commercial application of AI in home scenarios is gradually taking shape, bringing convenience and transforming everyday life.

To meet the diverse needs of home scenarios, we have studied the planning of home AI systems. Household data, including documents, images, and videos, along with critical data types like household and user profiles, is a key assent. Building a home knowledge base around this data has become central in the development of home AI. An intelligent agent-based technical solution supports home AI services in both proactive and reactive modes. The home knowledge base and the home intelligent agent work in tandem to enable applications across various areas such as health, learning, entertainment, work, fitness, and smart living (Fig. 1).

Home Knowledge Base: The Core of Home AI Services

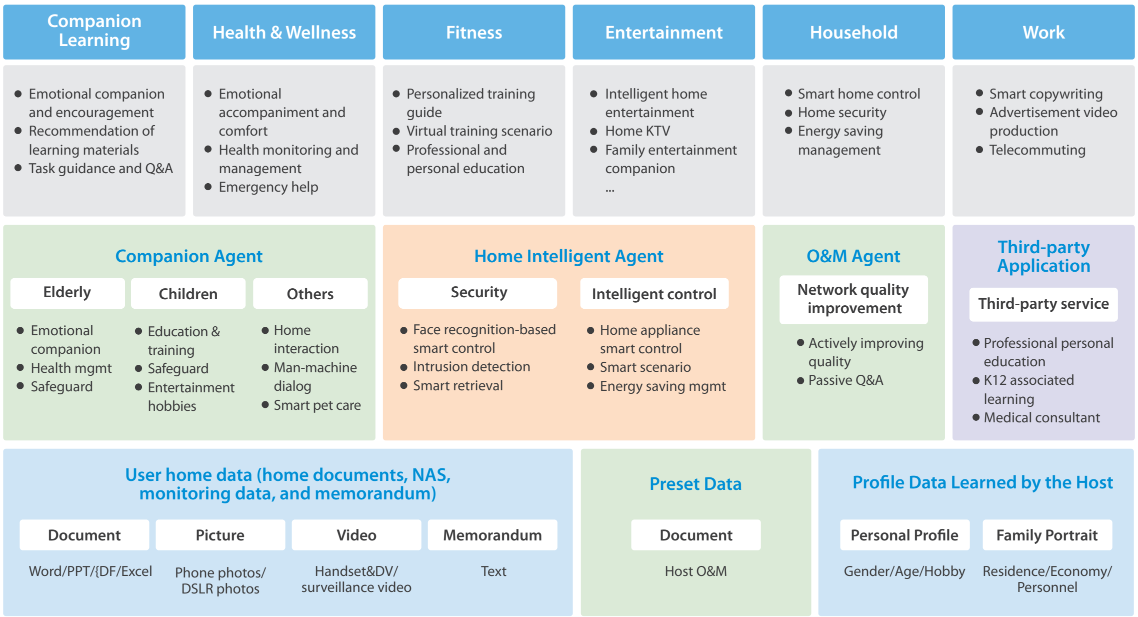
The home knowledge base is a knowledge system that integrates various domains, such as childcare, learning, health, work, entertainment, daily life, and fitness to provide scientific guidance and personalized services for family members. The data includes:

- **Home document data:** Includes documents uploaded by users, such as PPTs, Word files, PDFs, images, and videos, as well as user memos.
- **User & home profile data:** Gathers user behavior, preferences, health data, and home environmental conditions through sensors and interaction records to construct user and home profiles.
- **Home knowledge graph:** Uses knowledge extraction and structured representation to build a graph that maps relationships between family members, devices, and the environment.
- **Home device and environmental data:** Includes sensor data (e.g., temperature, humidity), device logs, environmental change records, and basic information, usage history, and maintenance records of smart devices.
- **Other home knowledge bases:** Contains general knowledge bases, such as encyclopedic knowledge graphs uploaded by users, and curated data.

Home Intelligent Agent: The Pathway to Home AI Services

The home intelligent agent leverages data from the home knowledge base and employs machine learning, natural language processing, and multimodal large models to understand family members' needs and deliver personalized services. Its key functions include:

- **User profile construction:** Builds detailed profiles for each family member, covering habits, preferences, and behavioral patterns, by combining user behavior data from the home knowledge base with family member information and preference settings.
- **Scenario awareness and dynamic adjustment:** Perceives changes in the home environment and provides scenario-appropriate services by integrating environmental data and device status from the home knowledge base to identify the scenario.
- **Multimodal interaction and natural language understanding:** Supports user interaction through voice, text, gestures, and other modalities, understanding complex instructions and using contextual and historical data to



Companion Learning
- Emotional companion and encouragement
- Recommendation of learning materials
- Task guidance and Q&A

Health & Wellness
- Emotional accompaniment and comfort
- Health monitoring and management
- Emergency help

Fitness
- Personalized training guide
- Virtual training scenario
- Professional and personal education

Entertainment
- Intelligent home entertainment
- Home KTV
- Family entertainment companion

...

Household
- Smart home control
- Home security
- Energy saving management

Work
- Smart copywriting
- Advertisement video production
- Telecommuting

Companion Agent

Elderly
- Emotional companion
- Health mgmt
- Safeguard

Children
- Education & training
- Safeguard
- Entertainment hobbies

Others
- Home interaction
- Man-machine dialog
- Smart pet care

Home Intelligent Agent

Security
- Face recognition-based smart control
- Intrusion detection
- Smart retrieval

Intelligent control
- Home appliance smart control
- Smart scenario
- Energy saving mgmt

O&M Agent

Network quality improvement
- Actively improving quality
- Passive Q&A

Third-party Application

Third-party service
- Professional personal education
- K12 associated learning
- Medical consultant

User home data (home documents, NAS, monitoring data, and memorandum)

Document	Picture	Video	Memorandum
Word/PPT/{DF/Excel	Phone photos/ DSLR photos	Handset&DV/ surveillance video	Text

Preset Data

Document
Host O&M

Profile Data Learned by the Host

Personal Profile	Family Portrait
Gender/Age/Hobby	Residence/Economy/ Personnel

Fig. 1 Implementing home AI services across various scenarios based on the home knowledge base and intelligent agent.

provide more accurate services.

- **Active learning and feedback:** Continuously learns and adapts to improve service quality based on family members' needs, collecting feedback to dynamically adjust behavioral strategies.
- **Personalized recommendations and proactive services:** Predicts family members' needs based on user profiles and scenario awareness, offering personalized services proactively.
- **Privacy protection and security:** Ensures compliance with privacy and security requirements for the home knowledge base and agent. Data is stored locally to prevent leakage, encryption protects sensitive information, and a user permission management system is in place.

Scenario-Based Implementation

Home AI services are tailored to specific scenarios to meet diverse needs:

- **Health and wellness:** Focuses on health monitoring and safety protection, with future developments emphasizing proactive intervention, such as automatically triggering emergency responses upon detecting abnormal behavior in elderly family members.
- **Companion learning:** Currently includes educational robots and smart learning platforms, with future advancements integrating multimodal interaction technologies for emotional interaction and adaptive teaching.
- **Entertainment:** Currently involves voice control and multi-device coordination (e.g., smart speakers controlling lights, music, and TV), with future trends in holography, socialization, and multi-user virtual engagement.
- **Work:** Current smart home features optimize the environment and efficiency, with future developments extending to intelligent collaboration.
- **Fitness:** Currently relies on the coordination of smart wearables and home fitness equipment, with future innovations integrating biometric recognition and virtual coaching, such as personalized training plans.
- **Smart home:** Currently achieves device interconnectivity and automated control, with future directions focusing on whole-house intelligence and energy self-sufficiency, such as integrating household energy for optimized allocation.

By leveraging the home knowledge base and intelligent agent, AI is evolving from single-function applications to full-scenario services, shifting household services from passive response to active care and driving improved quality of life. ZTE TECHNOLOGIES

Shandong Mobile and ZTE Harness Multi-Agent Collaboration for End-to-End Fault Management

China Mobile serves over one billion subscribers and operates the world's largest 5G network. As the leader in the modern mobile industry chain, it spearheads national tech innovation and drives coordinated ecosystem growth. Autonomous network (AN) is the first of 10 key sub-chains launched by the group, aimed at lifting O&M efficiency through AI and accelerating the journey to high-level autonomy.

In 2024, China Mobile issued joint research propositions for the AN sub-chain to its provincial branches and industry partners, calling on the industry to focus on cutting-edge AI technologies and join efforts to overcome bottlenecks in AN.

The rapidly increasing network scale and business complexity, coupled with scattered O&M tools and know-how, are throttling fault response. Faster detection, accurate root cause localization, streamlined dispatch, and closed-loop handling—while easing engineer workloads—has become mission-critical.

In response, Shandong Mobile and ZTE have partnered under China Mobile's AN sub-chain, forming a joint team harnessing AI O&M large models to drive breakthroughs in fault-monitoring.

Integrated AI Computing Appliance: A Foundation for Large-Model Innovation

GPU shortage poses a major risk to large-model deployment. To address this, the project team carried out in-depth investigations into networking conditions and site surveys. After extensive discussions on solutions, a full-stack intelligent computing solution was developed—taking only 45 days from demand initiation to implementation—thus laying a solid foundation for the innovative application of large models.

ZTE provides a full-stack intelligent computing

Zhou Bo

Senior System Engineer, Service Products, ZTE

Song Peishuang

Technical Support Expert, ZTE

solution covering computing power, network, capabilities, intelligence, and applications, meeting the differentiated requirements for performance, cost and service requirements across diverse scenarios. The appliance integrates high-performance GPUs, user-friendly training and inference platforms, and mainstream large models, solving the "last mile" problem in the commercial implementation of large models.

Through hardware-software collaborative optimization, GPU performance is maximized, while development complexity is reduced. End-to-end toolchains covering data preparation, model training, and inference, significantly lower the technical barriers for enterprises to develop AI applications, enabling operators to go from idea to production within hours.

The appliance supports ZTE's Nebula Telecom Large Model and mainstream open-source large models, offers one-click RAG deployment, and secures the entire pipeline with full-link encryption and zero-trust identity controls—delivering fast, secure, and stable inference while ensuring data security and privacy for operators.

Exploring "AI+" Fault Monitoring Innovation

Powered by the Nebula Telecom Large Model engine, ZTE has introduced large–small model collaboration and multi-agent collaboration to empower "AI+" fault monitoring scenarios. This enables more accurate fault identification and root-cause analysis, more efficient process integration for faster fault closure, and smarter intent-driven O&M, reducing the workload of O&M engineers.

Multi-Agent Collaboration Drives Efficient Closed Loops

As shown in Fig. 1, the solution adopts a two-level intent routing strategy. The master control agent performs intent recognition, routing, and process control, assigning tasks to specialized business agents such as identification, analysis, scheduling, and evaluation. These business agents complete their respective tasks and return results to the master

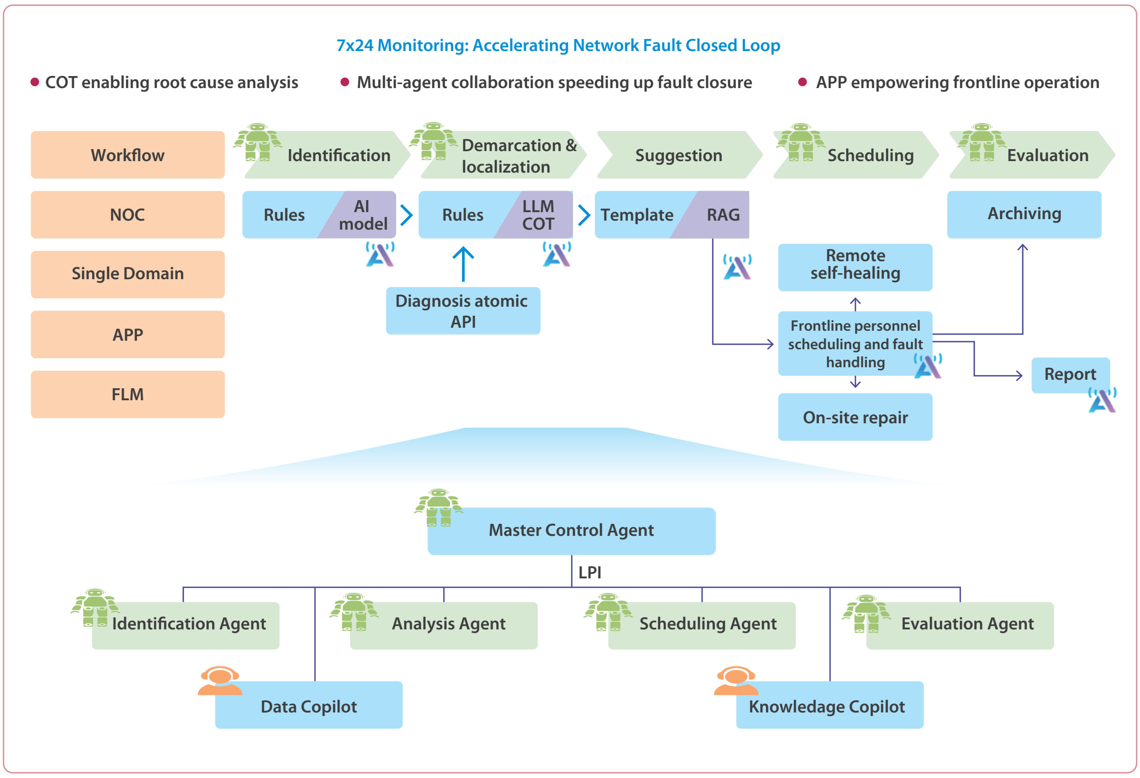


Fig. 1 Multi-agent collaboration for network fault monitoring.

> Powered by the Nebula Telecom Large Model engine, ZTE has introduced large–small model collaboration and multi-agent collaboration to empower "AI+" fault monitoring scenarios.

control agent, which then determines whether to proceed to the next stage. Through such collaboration, the fault handling process is driven towards a closed loop.

To minimize error accumulation in agent collaboration, the paradigm is shifting from API to language programming interface (LPI). By enabling agents to interact through LPI, the accuracy of multi-agent collaboration is enhanced.

Large–Small Model Collaboration Enhances Event Detection

In the fault detection phase, an identification agent is established leveraging large–small model collaboration technology. The small model handles dynamic data aggregation, while the large model generates event summaries, providing concise yet comprehensive event conclusions and enabling intelligent event generation within one minute.

CoT Reasoning Reshapes Fault Localization Capabilities

In the fault analysis phase, an innovative fusion of fault knowledge and chain-of-thought (CoT) causal reasoning is introduced. This approach performs comprehensive reasoning based on the fault case database and alarm data, improving the accuracy of cross-domain fault analysis across multiple factors to over 91%.

Shifting Backend Capabilities to Mobile

A scheduling agent and a knowledge copilot are developed based on large models, bringing fault knowledge, operational data, and atomic API capabilities to the mobile app. This enhances the on-site engineer's ability to solve problems independently and improves the interaction efficiency between frontline and back-office teams.

Decoupled Capabilities Accelerate Value Creation

By opening up capabilities and embedding AI into the existing network fault management system, scheduling system, and mobile app, pilot verification has been carried out in cross-domain fault scenarios spanning IP networks, transmission networks, and power and environment systems—achieving 1-minute intelligent fault detection, 91% root cause analysis accuracy, and minimized human effort.

This project practice has been recognized by TM Forum’s GenAI IG1345, the China Communications Standards Association (CCSA), ICT China 2024 Excellent Cases, as well as receiving the 2024 BRICS Industrial Innovation Contest Excellent Project Award, providing a valuable reference for large-model applications in the telecommunications industry.

In the future, ZTE and Shandong Mobile will continue to deepen their cooperation, further expand value-driven fault monitoring scenarios, accelerate the integration of AI innovation into O&M workflows, reduce the workload of O&M personnel, enhance operational efficiency, and achieve a closed loop of value creation and results. ZTE TECHNOLOGIES

ZTE's Wireless LLM Multi-Agent Assurance Solution Ensures Network Stability During Chinese New Year

In the era of rapid advancements in global communication technology, intelligent operations and maintenance (O&M) has become a central driver for industry transformation. As a leading player in the telecom sector, ZTE has been promoting industrial change through innovative technologies. At the 10th World Internet Conference (WIC) Wuzhen Summit in 2023, ZTE unveiled the first wireless large language model (LLM) agent, marking a new milestone in the intelligent O&M of communications networks.

After more than a year of technological iteration and practice, ZTE's wireless LLM has evolved from a single-agent architecture to a multi-agent intelligent center. This upgrade not only optimizes the technical architecture, but also enables more efficient and intelligent coordinated network operations.

The multi-agent intelligent center allows the system to invoke multiple agents on demand—like an intelligent corps with a clear division of labor and coordinated operations—greatly improving the efficiency and precision of network O&M.

To date, ZTE's wireless LLM application has been successfully deployed in 23 projects across China, spanning provinces such as Zhejiang, Anhui, Yunnan, Guangdong, Jilin, Xinjiang, Hubei, Chongqing, Hainan, and Fujian. In collaboration with local operators, ZTE has successfully implemented the model in real-world O&M scenarios, accelerating the intelligent and automated transformation of network management.

During the critical 2025 Spring Festival guarantee, ZTE's three major O&M agents were applied for the first time, successfully overcoming the challenges posed by the broad scope and extended duration. This highlighted the powerful advantages of multi-agent collaborative operations.

Throughout the guarantee period, the network faced the dual challenges of high traffic and complex

Deployment of the large-model Network Assurance Expert by Wuhan Telecom at the Wuhan sub-venue of the 2025 Spring Festival Gala.

scenarios. ZTE's LLM multi-agent system managed the entire process, addressing these challenges through three key agents:

- **Network Insight Expert:** Acts as a "panoramic scanner", proactively identifying risks (e.g., weak coverage or low traffic) through big data analytics and AI algorithms, resolving issues before they escalate.
- **Network Assurance Expert:** Focuses on critical events with fully automated, end-to-end workflows. It dynamically allocates network resources to high-traffic areas (such as tourist attractions and transportation hubs) and continuously monitors user experience metrics to ensure stable service quality.
- **Fault Guard Expert:** Functions as a "crisis response unit", swiftly diagnosing major faults by integrating historical data with real-time insights. It pinpoints root causes and delivers precise solutions to minimize service disruptions.

In collaboration with Wuhan Telecom, ZTE provided assurance for 238 5G cells across key locations, including the Huanghelou Spring Festival Gala sub-venue, Optics Valley Spring Festival Gala sub-venue, Wuhan Railway Station, SKP mall, and Liyuan park at East Lake. Through LLM multi-agent coordination, the network ensured smooth live video streaming and live interactions.

The successful application of the LLM during the Spring Festival not only delivered all-round network assurance but also verified the effectiveness of the multi-agent solution through real-world data, improving user experience and customer satisfaction.

Looking ahead, ZTE will continue to embrace an innovation-driven growth philosophy and further optimize LLM intelligent agent technologies. From a technological perspective, ZTE will further explore the deep integration of LLM with communication networks, optimize collaborative agent algorithms, and enhance the efficiency of multi-agent collaboration in complex scenarios. Additionally, ZTE will strengthen the adaptability of LLM to new services, including 5G industry applications and emerging 6G technologies. On the service side, ZTE will provide operators with tailored and more efficient O&M solutions, offering intelligent support throughout the entire lifecycle—from network planning and construction to operations and maintenance. By working closely with industry partners, ZTE aims to open a new chapter in the development of communications, ensuring that the benefits of intelligent O&M extend to all sectors of society. ZTE TECHNOLOGIES

Wang Hongxin

Senior Network Technology Engineer, ZTE

Wang Kang

Chief System Development Engineer for Network Technology, ZTE

Cao Lili

Senior Wireless Expert in Network Management Software Development, ZTE

ZTE

To lead in connectivity and intelligent computing, enabling communication and trust everywhere