GSMA | GTI

# Mobile AI

March 2026

# Contents





# Executive Summary

The digital economy has become a primary engine of global economic growth, driven by the rapid evolution of mobile communication technologies and the accelerating advancement of Artificial Intelligence (AI). As 5G reaches global scale, 5G-Advanced moves towards commercialization and 6G research advances steadily, mobile networks continue to expand the limits of connectivity, efficiency, and coverage. At the same time, AI is extending from the cloud to devices and the network edge, with Large Language Models (LLMs), multimodal intelligence, and AI agents reshaping how intelligence is created and consumed. The deep convergence of these two forces is giving rise to a new era – **the Mobile AI era**.

In the Mobile AI era, the ubiquitous connectivity of mobile communication networks becomes the essential foundation for the widespread adoption and democratization of AI. Built on this pervasive connectivity, AI is expanding beyond specialized domains to become universally accessible. At the same time, deep integration with AI is reshaping the core architecture of those networks. Together, these forces make the convergence of mobile communication and AI an inevitable direction for technological evolution and industrial transformation. As a result, **Mobile AI has emerged as a transformative paradigm and a critical driver of global digital and intelligent development**.

Mobile AI is founded on the principle of bidirectional empowerment between networks and AI, underpinned by the core values of Responsible AI and Security & Trustworthiness. By integrating the ubiquitous connectivity, low latency, and high reliability of mobile communication networks (5G/5G-A/6G) with AI's strengths in perception, decision-making, and learning, Mobile AI establishes a collaborative **device-edge-network-cloud architecture**. This forms an intelligent service system that offers broad coverage, real-time responsiveness, and precise adaptation, with the overarching goal of delivering **AI that is ubiquitous, trustworthy, and easy-to-use**.

The essence of Mobile AI is captured in its "**Three-Layer, Four-Dimension**" architecture. Vertically, the Foundation Layer, the Execution Layer, and the Application Layer form an end-to-end pathway from infrastructure to scenario-specific value creation. Horizontally, four key functional dimensions - **AI for Network, Network for AI, Mobile AI Agents/Terminals, and Mobile AI Applications-** work together to drive continuous evolution across the entire ecosystem. These layers and dimensions interact dynamically, enabling full-chain integration from technology to application and accelerating the intelligent transformation of industries.

Drawing on a global industry perspective, this white paper systematically outlines the value, definition, architecture, implementation pathways, and key challenges of Mobile AI. It provides strategic recommendations and calls for collaborative actions across the value chain to promote global cooperation, standard alignment, and ecosystem growth. Looking ahead, as 5G-A becomes widespread and 6G advances toward commercialization, Mobile AI will evolve from converged application to native symbiosis. This shift will unlock sustained innovation momentum, empower industries at scale, and serve as a cornerstone for the high-quality development of the global digital economy.

# 01

# The Value of Mobile AI

The current large-scale deployment of 5G/5G-A networks and the rapid proliferation of Artificial Intelligence (AI) are propelling mobile communications and AI into a new phase of deep synergy. Within this trend, Mobile AI has emerged. It is not a mere superposition of the two technologies but is fundamentally based on the core logic of "bidirectional empowerment between networks and AI." By realizing "networks enabling intelligence, and intelligence optimizing networks," it constructs a new type of service system characterized by ubiquitous connectivity, real-time responsiveness, and intelligent self-adaptation. This "bidirectional empowerment" fusion paradigm is systematically reshaping industrial development trajectories and societal operations. It not only opens up new growth avenues for operators and the industrial chain but also provides unprecedented foundational support and innovative possibilities for the refinement of social governance, the digital and intelligent transformation across all industries, and the intelligent enhancement of users' daily life experiences.

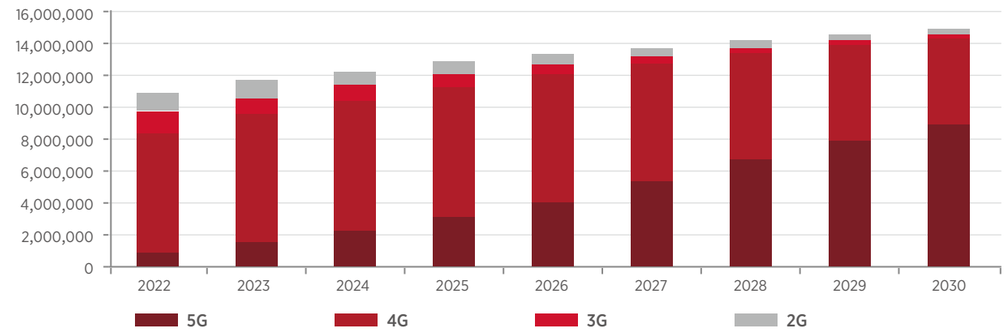
## 1. Economic Value: Unlocking Large-Scale Potential Through Convergence

The iterative upgrade of mobile communication technology and the pervasive penetration of AI are converging in a mutually reinforcing trend, establishing themselves as the core engine driving the development of the global digital economy. On one hand, the large-scale deployment of 5G networks and the technological evolution of 5G-Advanced (5G-A) provide a solid network foundation for AI applications. On the other hand, the rapid iteration and market expansion of AI technology impose higher demands on communication network capabilities. The deep integration of the two is breaking down industrial boundaries, catalyzing novel application scenarios and business models, and activating immense global market potential.

### 1.1 Network Foundation: Scale Network Deployment Drives Economic Contribution

The global mobile communication industry has entered a critical stage of 5G scale development. The widespread adoption of 5G technology has established an extensive network foundation for its convergence with AI. Since the first commercial 5G deployments in 2019, a total of 384 commercial 5G networks had been launched globally by the end of 2025, with the number of 5G subscribers surpassing 3 billion. From a long-term development perspective, 5G subscriber numbers are expected to maintain rapid growth. Forecasts indicate that by 2028, 5G will surpass

**Global Mobile Subscription Forecast (in 000s)**




Source: Omdia



Figure 1： Global Mobile Subscriber Forecast

4G to become the dominant global mobile communication technology. By 2030, global 5G connections are projected to reach 8.8 billion, accounting for over 60% of total global mobile connections, forming an extensive and densely connected network ecosystem.

By the end of 2025

**384**
Commercial 5G Networks

**3 billion**
5G Subscribers Surpassing

By 2030

**8.8 billion**
Global 5G Connections

**>60%**
Total Global Mobile Connections

According to Omdia predictions, the world's first 6G networks are expected to commence deployment and operation in 2029. By 2030, global 6G device connections will reach 289 million, growing to 3.5 billion by 2035, representing 22.3% of total global mobile device connections. Accompanying global RAN infrastructure investment is projected to rise sharply from $40 billion in 2030 to $250 billion in 2035, reserving next-generation network capabilities for Mobile AI's technological evolution. The mobile communications industry has already become a central pillar of global economic growth.

The GSMA Mobile Economy 2025 report indicates that in 2024, global mobile technologies and services generated $6.5 trillion in economic value added, representing 5.8% of global GDP. This contribution stems from a productivity boost effect exceeding $4 trillion and a direct mobile ecosystem contribution of $1.6 trillion, forming a complete value loop when combined with indirect economic impacts. Looking forward, with the large-scale adoption and deep integration of digital technologies like 5G, IoT, and AI, the dividends of technological empowerment will continue to be unleashed, steadily enhancing global productivity and operational efficiency. It is projected that by 2030, the mobile communications industry's contribution to the global economy will increase to nearly $11 trillion, raising its share of global GDP to 8.4%, cementing its role as the core engine for high-quality global economic growth.

According to Omdia Predictions
Global 6G Device Connections

**$289 million** (2030) → **$3.5 billion** (2035)

According to Omdia Predictions
Global RAN Infrastructure Investment

**$40 billion** (2030) → **$250 billion** (2035)

## 1.2 Innovation Driver: AI Investment and Market Growth Accelerate Convergence

The industrial application of AI is entering an explosive period. Global AI investment continues to expand, market space is broadening, and its convergence with mobile communication technology further amplifies its commercial value, enabling exponential growth.

On the consumer side, the adoption of Generative AI (GenAI) is accelerating significantly. According to Omdia's 2025 global consumer survey data, approximately 75% of respondents already use GenAI applications in scenarios such as daily life, work, and social interactions, with 38% opting for paid premium versions, both figures showing marked increases from the previous year. Smartphones have become the primary vehicle for consumers to access GenAI services: 48% of respondents use mobile GenAI apps for text search, 14% engage in conversations with Mobile AI agents, while 7% and 5% respectively use mobile GenAI apps to generate images/ videos and upload them online. This widespread consumer adoption is spawning new service demands, with 29% of global respondents expressing interest in having GenAI application subscription services bundled into their mobile service plans, pointing to new business model directions for operators.

**48%** mobile GenAI apps for text search

**14%** conversations with mobile AI agents

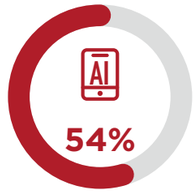
**7%** generate images and upload them online

**5%** generate videos and upload them online

The advancement of AI technology will also drive the proliferation of various new terminal devices. Omdia forecasts that by 2028, smartphones equipped with generative AI capabilities will constitute more than half (54%) of global smartphone shipments. Over the next decade, general-purpose embodied intelligent robots are poised for explosive growth, with global shipments projected to reach 2.6 million units by 2035, representing a compound annual growth rate (CAGR) of 85% over ten years.

According to Omdia Predictions

By 2028




smartphones equipped with generative AI capabilities

By 2035


compound annual growth rate (CAGR) over ten years.

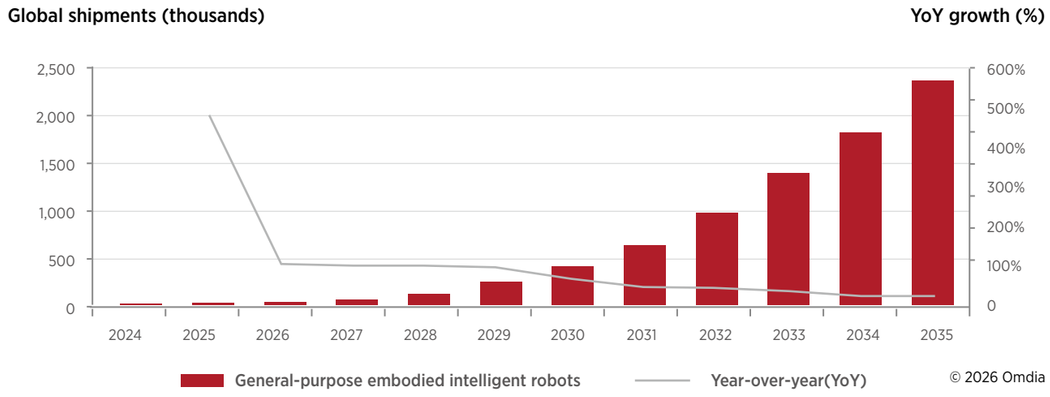



Figure 2: Global General-Purpose Embodied Intelligent Robot Shipment Forecast

On the enterprise side, the AI market is exhibiting explosive growth. Omdia predicts that by 2030, over 70% of the workload in global data centers will be dedicated to meeting AI-related computing demands. The GenAI software market is forecast to grow from $14.2 billion in 2024 to $101.3 billion in 2029, reaching $122.8 billion in 2030. As a core growth segment within the AI market, AI agent applications are developing even faster than the overall GenAI market. Their market size is expected to surge from $159 million in 2024 to $1.51 billion in 2025, potentially reaching $41.77 billion by 2030. Their share of the GenAI market is projected to rise from 6% in 2025 to 31% in 2030. Key enterprise AI agent use cases include automated code development/assistance, virtual assistants, intelligent process automation, and intelligent document processing.

According to a GSMA 2025 specialized survey involving 5,320 enterprises across 32 countries and 10 vertical industries, businesses plan to invest approximately 10% of their revenue into digital transformation initiatives, between 2025 and 2030. Investments in AI, along with 5G connectivity and terminal technology, rank at the top, serving as the core engine for the next phase of industry growth. Across all industries, over 90% of surveyed enterprises believe generative AI holds critical strategic significance for their digital transformation journey

**The GenAI software market is forecast**

**$14.2 billion** 2024 → **$101.3 billion** 2029 → **$122.8 billion** 2030

**AI agent applications market size**

**$159 million** 2024 → **$1.51 billion** 2025 → **$41.77 billion** 2030

## 1.3 Operator Strategy: Advancing Mobile AI to Expand Business Value

The GSMA Mobile Economy 2025 report shows that global mobile operator revenue reached $1.08 trillion in 2024. While revenue growth faces pressure from intense market competition and moderate ARPU growth, a steady long-term upward trend is expected, with revenue projected to increase to $1.25 trillion by 2030. To support technological iteration and business expansion, total global mobile operator capital expenditure from 2024 to 2030 is estimated at $1.3 trillion. This investment will focus on unleashing 5G's potential across all scenarios and driving enterprise digital transformation, thereby solidifying the infrastructure foundation for Mobile AI's convergent development.

The widespread adoption of AI applications is directly driving explosive growth in related network traffic. Omdia predicts that the compound annual growth rate (CAGR) of global AI traffic will reach 73% during the period from 2025 to 2033, with the year 2031 marking the critical crossover point where AI network traffic surpasses traditional traffic. Concurrently, small, cost-effective open models, exemplified by DeepSeek, are propelling AI deployment towards edge devices. The CAGR for new AI traffic flowing to the network edge from 2025 to 2033 is projected to reach as high as 130%. This growth in traffic volume and the evolution of traffic characteristics not only drive

unleashing 5G's potential across all scenarios

**$1.3 trillion**
total global mobile operator capital expenditure from 2024 to 2030

driving enterprise digital transformation





**Total network traffic, AI and non-AI, 2023-33**

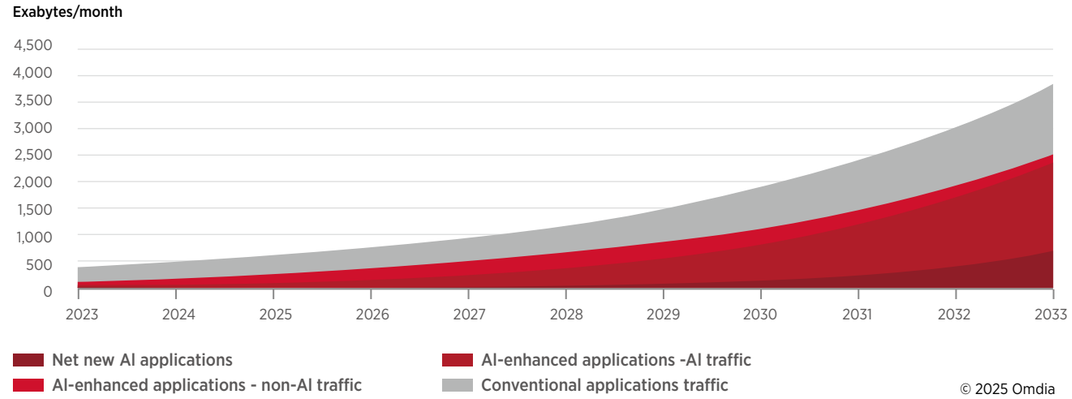





Figure 3: Network Traffic Forecast

telecommunications operators to continuously enhance network transmission capabilities but also trigger long-term optimization of network design principles.

Confronting the wave of AI technological development, global telecom operators are actively advancing strategic transformation. They are proactively addressing the traditional industry challenge of "increased traffic without proportionate revenue growth" by exploring the value potential of Mobile AI convergence across three core dimensions:

**Building Dedicated AI Infrastructure:** Upgrading broadband networks to accommodate AI service demands, constructing AI data centers to offer GPU-as-a-Service (GPUaaS), and exploring integrated layouts for computing power with metropolitan area networks and access networks to promote the scaled development of edge AI computing.

**Enhancing Own Operations with AI:** Deploying AI applications in areas such as network management, customer service, and personal assistants to achieve efficiency gains across the entire operational process.

**Directly Venturing into AI Software Services:** Launching generative AI-based products and services for consumers, enterprises, and vertical industries.

Omdia's 2025 Global Operator Survey Report indicates that 40% of surveyed global operators have already deployed AI agent-based autonomous network solutions in commercial or trial networks, with an additional 34% planning to initiate related work within the next two years. This proactive participation by operators is accelerating the industrialization of mobile communication and AI convergence.

The integration of mobile communications and AI has moved from technical exploration into the phase of scaled implementation. Supported by 5G and 5G-A networks, and amplified by the synergistic effects of the AI market boom and continuous computing power expansion, it provides diversified support for the digital and intelligent transformation of entire industries. This drives industry development to leap from "localized optimization" to "systemic reconstruction," achieving a core advancement from cost reduction and efficiency gains to value creation. The deep fusion of the two continues to broaden industrial value boundaries, comprehensively enabling the digital and intelligent upgrade of all industries, from smart manufacturing to smart cities. Trillion-dollar market potential is accelerating its release. Over the next five years, the integration of mobile communications and AI is expected to deepen continuously. By the time of global 6G commercial deployment around 2030, the industry will enter an entirely new development stage characterized by the comprehensive native fusion of mobile services and AI capabilities.

## 2. Economic Value: Unlocking Large-Scale Potential Through Convergence

Leveraging its technical characteristics of ubiquitous connectivity and intelligent collaboration, Mobile AI deeply integrates into the core facets of societal operation. Its social value is primarily manifested in three key dimensions: first, empowering the transition of social governance towards greater refinement and efficiency; second, promoting the breakdown of barriers to achieve universal access to AI technology (AI democratization); and third, facilitating AI compliance and safety governance through real-time connectivity and remote controllability. Together, these three dimensions work synergistically to build a fair, inclusive, reliable, and sustainable social development ecosystem, injecting enduring momentum into societal progress in the digital economy era.

### Enabling Intelligent Upgrading of Social Governance

Building on the combined capabilities of "Communications + AI," Mobile AI constructs a governance system featuring holistic perception, rapid response, and precise intervention. This drives the transformation of governance models from "experience-driven" to "data-driven." In the domain of urban governance, it aggregates multidimensional data (such as municipal affairs, transportation, and environment) through holistic sensing and intelligent analysis. This enables dynamic early warning and targeted measures for pain points like traffic congestion or utility leaks, thereby enhancing operational efficiency, reducing governance costs, and advancing the evolution towards refined and intelligent management. In public services, analyzing supply and demand data helps accurately identify resource gaps, guiding high-quality resources to underserved areas and narrowing urban-rural and regional disparities. Functions like intelligent appointment scheduling and process optimization streamline procedures, improving service accessibility and public satisfaction. For emergency response, leveraging low-latency communication and intelligent analytics creates an integrated system capable of early identification of sudden risks, rapid warning, and efficient incident management. Cross-departmental coordination shortens response times and mitigates disaster losses.

### Promoting Universal Access and Democratization of AI Technology

Mobile AI leverages widely adopted intelligent mobile terminals to break down the hardware barriers, technical complexities, and scenario limitations associated with AI technology application. It brings professional-grade AI capabilities to all industries and households. This facilitates the transition of AI from an "exclusive technological highland" to a "universally accessible" resource. It plays a crucial role in scenarios such as personalized tutoring in rural education, assisted diagnosis in primary healthcare, intelligent services for inclusive finance, age-friendly intelligent adaptation for older adults, and digital empowerment in remote areas. It significantly reduces the application cost and usability threshold of AI technology, enabling equal access to the convenience and value brought by AI for people across different regions, demographics, and industries. This achieves comprehensive coverage and inclusive sharing of AI capabilities, ensuring the developmental dividends of intelligent technology benefit every individual.

### Empowering AI Compliance, Safety, and Controllability

Based on real-time, reliable mobile communication networks, Mobile AI provides remote supervision and safety takeover capabilities for advanced intelligent devices such as embodied intelligence systems and autonomous vehicles. It builds a compliance and security assurance system spanning the entire AI lifecycle. At the device operation level, low-latency connections enable real-time status monitoring and intervention for abnormal behavior in intelligent terminals like autonomous vehicles and robots. In the event of sudden risks, remote safety modes can be activated or control can be taken over, greatly enhancing system fault tolerance and personal safety assurance in complex environments. At the ecosystem collaboration level, by connecting multiple regulatory bodies, enterprise technology platforms, and end-users, a dynamic, collaboratively responsive AI governance network is established. This promotes the formation of a closed-loop encompassing "R&D - Deployment - Operation - Supervision," ensuring the development of intelligent technology consistently operates within a safe, reliable, and controllable framework.



02

# The Connotation of Mobile AI

Mobile AI is fundamentally rooted in the core logic of "bidirectional empowerment between networks and AI," guided by the values of "Responsible AI" and "Security & Trustworthiness." It leverages the characteristics of advanced mobile communication networks (such as high bandwidth, low latency, high reliability, and ubiquitous connectivity) and combines them with AI's capabilities in intelligent perception, autonomous decision-making, and collaborative reasoning. Through the holistic synergy of "device-edge-network-cloud," it achieves hierarchical deployment and dynamic scheduling of intelligent capabilities, constructing a systematic "Three-Layer, Four-Dimension" architecture. Relying on its core features of bidirectional empowerment, holistic synergy, scenario-specific closed loops, and cyclical evolution, it enables full-chain integration and continuous iteration from the technological foundation to industrial value creation.

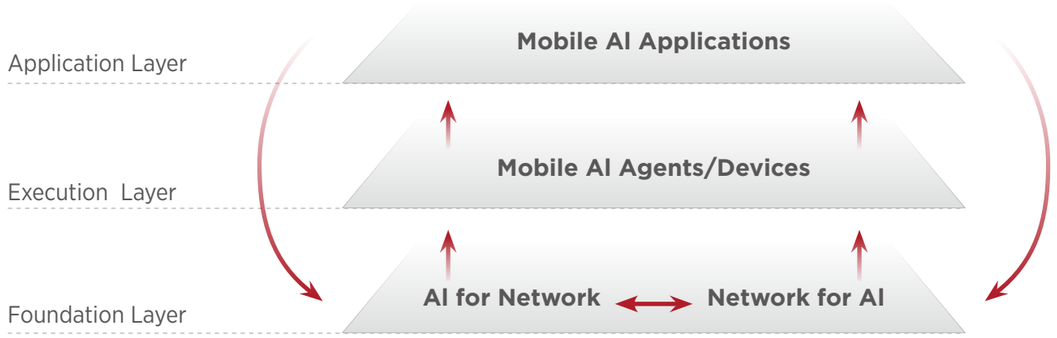



Figure 4: The Three-Layer, Four-Dimension Architecture of Mobile AI

The core architecture of Mobile AI manifests in the following three layers and four key dimensions. The "Three Layers" define the vertical implementation path from infrastructure to value creation, with each layer closely interconnected through capability supply and demand feedback:

**Foundation Layer:** Serving as the technological bedrock, it deeply integrates the bidirectional capabilities of "AI empowering networks" and "networks carrying intelligence," providing core support for holistic connectivity, heterogeneous computing power, and collaborative intelligence.

**Execution Layer:** Acting as the entity for executing and implementing capabilities, it is responsible for encapsulating the intelligent and connectivity capabilities from the foundation layer into deployable, interactive function execution units and providing standardized service interfaces to the application layer.

**Application Layer:** As the layer for value realization, it builds intelligent solutions within specific business scenarios based on services provided by the execution layer, directly generating economic and social benefits.

The "Four Dimensions" permeate and drive the operation of the entire system, representing the key functional dimensions for Mobile AI to achieve its objectives:

AI for Network: Deeply applies artificial intelligence throughout the entire lifecycle of communication networks, enabling the automation and intelligence of network planning, operation, maintenance, optimization, and management, thereby enhancing network efficiency and autonomy levels.

**Network for AI:** Utilizes the high-performance, high-reliability, and low-latency capabilities of mobile communications to meet the differentiated needs of intelligent applications for data transmission, computing power collaboration, and ubiquitous access, providing ubiquitous connectivity support for intelligence.

**Mobile AI Agents and Devices:** Asthe entities for large-scale hosting and interaction of intelligent capabilities, they encompass diverse forms such as robots, AI phones, wearable devices, various IoT terminals and agents. Through device-edge-cloud collaborative computing, they overcome the resource constraints of single devices, enabling the democratized deployment and personalized delivery of intelligent services.

**Mobile AI Applications:** Transform the converged technological capabilities into scenario-specific solutions for all industries and the consumer domain. Spanning key fields such as smart manufacturing, intelligent transportation, precision healthcare, and urban governance, they drive industrial upgrading and enhance societal operational efficiency.

These four dimensions form a dynamic cycle within the "Three-Layer" architecture: the continuous optimization of the foundation layer promotes innovation and capability upgrades in the execution layer, which in turn catalyzes scenario-specific innovation in the application layer. Conversely, the actual demands from the application layer drive the iteration of foundational technologies and the advancement of execution capabilities in reverse. This constitutes a closed-loop evolutionary system of "technology - execution - application," ultimately supporting Mobile AI in realizing its vision of making "intelligence ubiquitous, trustworthy, and user-friendly."

## 1. AI for Network

AI for Network refers to the deep application of Artificial Intelligence technology throughout the entire lifecycle of communication networks, encompassing planning, maintenance, optimization, and operations. Its goal is to enhance network performance, user experience, and operational efficiency, ultimately leading to network autonomy. Mobile AI will transform the current landscape of "add-on" and "fragmented" network intelligence solutions by deeply integrating AI technology into the very architecture of communication networks. This builds capabilities for autonomous network sensing, intelligent decision-making, and closed-loop optimization, making AI an inherent, native, and ubiquitous core function of mobile communication networks, thereby driving networks towards higher orders of intelligence.

### 1.1 Network Planning Based on Mobile AI

The surge in data traffic in the digital economy presents challenges for network planning, such as congestion, capacity management, and experience assurance. While AI solutions in the 5G era could predict capacity demands and optimize investments using historical data, their core reliance on past data resulted in weak dynamic adaptation capabilities. Faced with sudden traffic surges from new services or construction obstacles, repeated surveys and adjustments were often necessary, leading to prolonged cycles and inefficiency.

Mobile AI drives the transformation of network planning from "static design" to "dynamic growth," achieving autonomy, precision, and efficiency. On one hand, it integrates multi-domain data under privacy compliance to generate a real-time, panoramic view of the network, designs compliant plans, and simulates risks in extreme scenarios. On the other hand, it automates processes via standardized interfaces, deploys intelligent nodes to complete end-to-end operations like device onboarding, and uses devices like AI-enabled glasses to collect data in real-time and dynamically adjust plans, ensuring construction timelines are met. By enabling data synergy, dynamic execution, and real-time closed loops, Mobile AI addresses the pain points of traditional planning and construction, providing core support for the foundational networks of the digital society.

### 1.2 Network Operation and Maintenance Based on Mobile AI

With increasing network complexity and escalating O&M demands, traditional models are becoming unsustainable. As network scale expands, devices proliferate, and extreme weather events occur, manual operations are time-consuming, difficult, and prone to impacting service continuity. AI solutions in the 5G era, while capable of predicting failures of individual devices, relied on historical data, lacked the ability to diagnose unknown or complex faults, and required

manual intervention for repairs. This led to high response delays and potential errors. The independent deployment of AI capabilities across different domains also made collaborative handling of cross-domain faults challenging.

Mobile AI promotes the evolution of O&M towards efficient autonomy through continuous learning and closed-loop coordination. On one hand, it enables continuous learning and knowledge sharing, where experience from handling a novel fault in one area can be synchronized across the entire network. Concurrently, it dynamically constructs causal reasoning graphs to proactively predict risks and take preemptive actions. On the other hand, it creates a "perception-decision-execution-feedback" closed loop that automatically completes tasks like metric collection, anomaly identification, strategy generation, and device orchestration. It also facilitates cross-domain coordination to handle cross-domain faults collaboratively and dynamically schedule resources across the entire network. By endowing the network with autonomous thinking and action capabilities, Mobile AI drives the network towards self-optimization and self-healing, advancing the goal of full-domain autonomy and injecting core momentum for stable network operation and efficient iteration.

## 1.3 Network Optimization Based on Mobile AI

The continuous expansion of network scale and service types highlights the contradiction between static rules and dynamic traffic. Current AI optimization solutions are mostly "add-on" tools that rely on offline training to output fixed strategies. Model updates lag behind network changes, and execution across separate domains lacks coordination, making it difficult to achieve global optimization or precisely meet the differentiated needs of users.

Mobile AI reshapes the network optimization system, upgrading point-based offline algorithms into a systematic online capability. Through a cross-domain coordination mechanism, intelligent agents across domains dynamically negotiate resources and strategies to achieve end-to-end multi-objective optimization. Possessing intent comprehension capabilities, it translates business requirements into precise optimization strategies tailored to the core demands of different scenarios. It incorporates an online incremental learning module to collect data in real-time and fine-tune models, enabling dynamic, context-specific adaptation ("one policy per moment"). Combined with cloud-edge-device collaborative reasoning, this reduces response latency. Mobile AI drives network optimization towards "autonomous, precise, and efficient" transformation. It not only automatically maintains optimal performance but also proactively predicts and adjusts, providing users with high-quality, refined network assurance and solidifying the foundation of the digital service experience.

## 1.4 Network Operations Based on Mobile AI

Currently, AI technology is deeply integrated into all aspects of network operations, from content recommendation and service activation to user service and security protection, forming a relatively comprehensive auxiliary system. However, existing AI-assisted operational systems are mostly designed independently for specific scenarios, such as customer service chatbots, recommendation engines, and security monitoring systems. These systems often operate based on isolated datasets, lacking interoperability and collaboration across systems. This leads to delayed operational responses, which not only reduces operational efficiency but may also cause misjudgments due to one-sided decision-making data.

Mobile AI builds a collaborative intelligent operations system, breaking down silos between systems to enable cross-module interoperability and cooperation. By collecting and analyzing multi-domain data from the user side, network side, and business side, it constructs dynamic user profiles to accurately push personalized service plans and value-added services, thereby enhancing user experience and value. Simultaneously, it monitors network and service status in real-time, proactively identifies hidden experience issues, and adjusts operational strategies in advance to reduce user complaints while safeguarding user privacy and asset security. Mobile AI will establish a layered operational architecture that organically unifies real-time business processing, cross-domain task coordination, and strategic planning. This promotes the deep integration of human expertise and machine intelligence, offering a new pathway for innovation and upgrading in network operations.





## 1.5 Summary

Mobile AI technology can effectively address the pain points in current network planning, construction, maintenance, optimization, and operations, delivering substantial commercial value for operators.

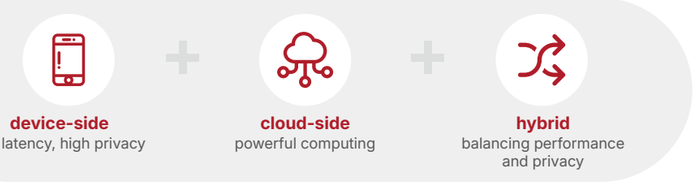
precise capacity planning and resource scheduling can prevent over-provisioning of network resources, directly saving energy and hardware investment costs. Predictive maintenance can also avert massive losses from major failures, achieving cost reduction and efficiency gains.

network-side refined service plan design and operations based on dimensions such as user, service, time, and location provide an excellent user experience, enhancing user retention, willingness to pay, and brand reputation. This empowers new services and drives business revenue growth.


network operations generate vast amounts of data. Under the premise of legality, compliance, and anonymization, AI technology can be used to mine value from this data. It provides data insights for product design, marketing, and network investment, and can also offer high-value data products to vertical industries, creating new revenue streams and enabling data monetization.

# 2. Network for AI

The deep integration of AI and communication networks gives rise to Mobile AI, driving the rapid emergence of new terminals such as intelligent robots, AI phones, and AI glasses. These devices are no longer mere tools but intelligent carriers with interaction and decision-making capabilities. Their large-scale development imposes higher demands on network connectivity services. Embodied intelligent robots, which achieve a "perception-decision-action" closed loop, have stringent requirements for network rate and latency for real-time interactive services (e.g., voice Q&A, environmental analysis) and remote teleoperation services. Remote teleoperation additionally demands mobility and high reliability. At the smart terminal level, AI phones adapt to different scenarios through three approaches: device-side (low latency, high privacy), cloud-side (powerful computing), and hybrid (balancing performance and privacy). AI glasses rely on the transmission of multimodal data like HD video and images, imposing high requirements on network latency, rate, and stability. The development of these new terminals and services necessitates that mobile communication networks like 5G-A specifically enhance their connectivity service capabilities to match the diverse and demanding network requirements.



**Accurately Identifying Multimodal Service Characteristics**

Mobile AI services are not monolithic, they encompass diverse scenarios such as data-intensive, latency-sensitive, and control-precision types. The core of connectivity assurance lies in the distinct multimodal characteristics of Mobile AI services. This means the network must simultaneously carry and guarantee the transmission of services with differing characteristics. Each service has its unique "modal fingerprint" and varying demands on network resources. This breaks the traditional network's predominant "best-effort" service model, requiring the network to possess modal-level service awareness and sophisticated differentiated service capabilities.

**Simultaneously Guaranteeing Multi-dimensional Network Performance**

Mobile AI services require networks to move beyond the traditional model of optimizing single metrics. They must achieve coordinated, multi-dimensional assurance on the same physical infrastructure, attaining multi-objective optimization for rate, low latency, and capacity under finite resource constraints. This necessitates flexible scheduling capabilities to dynamically allocate uplink/downlink resources and control latency jitter. Latency must be guaranteed based on service priority tiers: core control and real-time decision scenarios require ultra-low latency to ensure rapid command response; real-time interactive scenarios need low latency for a smooth experience; non-real-time data synchronization scenarios should maintain latency within a reasonable range. Bandwidth must elastically adapt to requirements like multimodal transmission and multi-device collaboration, catering to both lightweight interaction and high-bandwidth demands in heavy-load scenarios, and addressing the challenge of high-concurrency transmission from densely collaborating devices. Rate must balance stability and efficiency to ensure core business operations proceed normally, adapting to the differentiated needs of various scenarios.

**Achieving Cross-layer Data Collaboration and Sharing**

A collaborative barrier of "speaking different languages" and "being out of sync" currently exists between networks and Mobile AI services. As the underlying pipeline, networks typically provide "best-effort" services in a standardized, generic manner, lacking deep awareness of the dynamic needs of services. Conversely, service applications cannot obtain real-time network status (e.g., congestion, latency) and thus struggle to adaptively adjust their data transmission strategies. This necessitates that networks break away from the traditional "pipeline" transmission model and collaborate in real-time with services and terminals to achieve real-time cross-layer, cross-domain status awareness, resource scheduling, and coordinated control.

Mobile communication networks like 5G-A focus on "Enhanced Connectivity, Granular Assurance, and Cross-layer Collaboration." Through comprehensive upgrades in these three capabilities, they provide full-link support for new AI-powered intelligent services, terminals, and applications, propelling AI technology and applications from "theoretically feasible" to "large-scale deployment."

## 2.1 Enhanced Connectivity

Multi-dimensional Connectivity Enhancement for Improved Service Experience. Moving beyond the limitations of single-metric optimization in traditional networks, mobile communication networks like 5G-A holistically coordinate rate, low latency, capacity, and reliability. This supports the concurrent operation of multiple types of Mobile AI services on the same physical infrastructure, ensuring a high-quality, "seamless" connection experience. Uplink enhancement is achieved through technologies like Uplink Carrier Aggregation (CA) and Supplemental Uplink (SUL). The former bundles multi-frequency carriers to dynamically allocate data flows, while the latter deploys a dedicated uplink carrier in lower frequency bands. Both can provide ultra-large bandwidth and low latency, supporting the real-time transmission of HD video, sensor data, and control commands from intelligent devices. Intelligent Pre-scheduling learns service transmission patterns and, by evaluating multiple factors like channel quality, intelligently allocates uplink resources. This eliminates the waiting time for scheduling requests and grants, reducing data transmission latency.

## 2.2 Cross-layer Collaboration

Cross-layer Collaborative Transmission Across Network, Service, and Device. Mobile communication networks like 5G-A, leveraging modal-level service awareness and fine-grained differentiated service capabilities, can achieve "on-demand allocation and precise adaptation" of network resources for Mobile AI intelligent services. This addresses the service adaptation shortcomings caused by the traditional network's "one-size-fits-all" approach. Whether for data-intensive AI large model training and AI Agent intent reasoning, latency-sensitive autonomous driving real-time decision-making and instantaneous data transmission for embodied robots, or control-precision collaborative operation and manipulation of embodied robots, each can receive dedicated network resource configurations. On the other hand, intelligent services have high demands on network uplink. When in scenarios with weak coverage or strong interference, uplink rates are severely limited. In such cases, network-device-cloud coordination is required to adjust the processing mechanism. When the network confirms that performance metrics like uplink rate cannot meet demands, the terminal and cloud can collaboratively adapt, shifting from a combined device-cloud AI processing mode to a device-side AI processing-dominated mode.

## 2.3 Granular Assurance

Differentiated Characteristic Assurance for Multimodal Interaction. By identifying the multiple parallel modal data streams within a service, mobile communication networks like 5G-A ensure the prioritized and complete delivery of critical modal data. This enables a transition from a "coarse-grained pipeline" to fine-grained, modal-level assurance, tackling the challenge of complex service characteristics. Modal types in new intelligent service flows include audio, video, control, and other data streams, each with different QoS requirements and varying pattern characteristics (e.g., data volume, importance, latency & bandwidth, reliability). The network configures different QoS requirements and allocates corresponding transmission resources for data streams of different importance and modalities. For example, remote operation data streams for embodied intelligence can be configured with high-reliability QoS to guarantee high-priority data transmission. Other data streams receive standard QoS, allowing for fault-tolerant transmission over the air interface while maintaining user experience, thereby improving transmission performance and ultimately enhancing service quality and the effectiveness of wireless transmission.

Mobile AI service scenarios place demands on the network's "depth of connectivity, precision of collaboration, and granularity of assurance" that far exceed those of traditional consumer services. Networks must solve the adaptation challenges between service demands and network supply, realizing a closed loop of "service demand - network capability - service quality." This provides a solid digital foundation for the large-scale deployment of Mobile AI. In the future, device-side AI will rely on local computing power on phones, wearables, vehicles, etc., to deliver low-latency, high-privacy, offline-capable processing, interaction, and IoT coordination capabilities, ensuring real-time performance and personalization. The network, as the capability support hub, will evolve from a passive pipeline to an intelligent core, matching differentiated optimal assurance plans for different customers based on factors like network energy efficiency, latency, rate, and data priority. Cloud-side AI will support complex tasks like large model training and global reasoning with massive computing power and big data. Terminals, networks, and the cloud will together constitute the collaborative core foundation of the intelligent era.

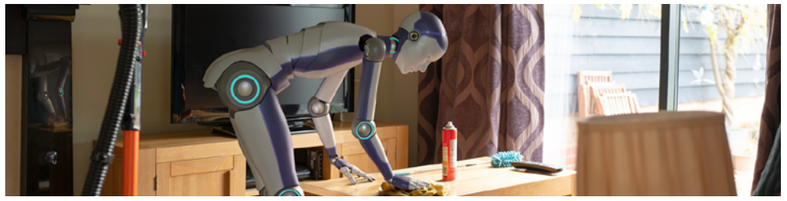
# 3. Mobile AI Agents/Devices

Mobile AI has given rise to diverse forms of intelligent terminals, such as embodied intelligent devices, AI agents, and wearable devices. These Mobile AI agents and devices are reshaping the modes of interaction between humans and devices, as well as between devices and their environments. Intelligence is no longer an isolated application dependent on the computing power of a single terminal but has evolved into a persistent capability that operates through collaboration across devices, networks, and computing domains. In this context, Mobile AI agents and devices face key implementation challenges: embodied intelligent robot terminals deeply embed AI into the physical world, demanding extremely high real-time performance, reliability, and security; Mobile AI smart terminals are constrained by computing power, power consumption, and cost, requiring a balance among multiple demands; terminal AI agents, evolving into system-level gateways, must deliver full-scenario services under resource constraints. This chapter focuses on two main directions: Mobile AI embodied intelligence, and smartphones, wearable devices, and agents. By combining industrial-grade and consumer-grade practical case studies, it systematically analyzes the key architectures and core capabilities required for their implementation, revealing their common requirements and differentiated value across different application.

## 3.1 Mobile AI Embodied Intelligence

### ❶ Embodied Intelligent Home Robots

terminal
edge
cloud



Home intelligent robots are entering a new stage of multi-scenario intelligent collaboration. They have evolved from single-function execution to intelligent cooperation, breaking the limitations of structured environments. Their potential for large-scale application in areas such as community services and home companionship is becoming prominent. Current trends are manifested in three main aspects: diversification of form factors, multi-robot collaboration, and modularization/standardization. However, core challenges remain: adaptation to complex environments, generalization of tasks across different scenarios, and control of deployment and operational costs.

Mobile AI addresses these core challenges with a "terminal-edge-cloud" collaborative hybrid architecture, building a multi-tiered capability distribution system. It employs a "centralized intelligence with distributed capabilities" design. A foundation large model handles complex task perception, while lightweight models on the terminal side enable real-time environmental awareness and basic decision-making, ensuring low-latency response. The edge layer undertakes regional collaborative computing and multi-robot collaboration scheduling. The cloud side provides large-scale training, cross-scenario knowledge base updates, and global resource optimization. The mobile communication network supports multimodal interaction and remote control, helping robots transcend scenario boundaries. This solution achieves a unified value proposition of adaptability, collaboration, and cost-effectiveness. It significantly enhances the robot's ability to adapt to complex environments, supports cross-scenario task generalization and multi-role collaboration. The parallel development of modular hardware and functional software drastically reduces deployment and operational costs. This promotes the large-scale adoption of home robots in scenarios such as domestic services and elderly care/rehabilitation, accelerates the penetration of AI technology into everyday life scenarios, and contributes to building an inclusive intelligent living ecosystem.

### ❷ Embodied Intelligent Terminal Factory Collaboration

device
edge

Intelligent factory multi-robot collaboration must balance real-time performance and reliability. Scenarios like collaborative transport of large components impose stringent requirements on real-time robot control, precise coordination, and safety assurance, relying on Cyber-Physical System scheduling. Traditional centralized cloud processing struggles to meet demands for millisecond-level response, low bandwidth consumption, and high-concurrency control. Compounded by complex factory environments and constraints on terminal computing power and energy consumption, deploying AI models and performing real-time inference becomes increasingly difficult, forming a key bottleneck for smart factory upgrades.

Mobile AI tackles the pain points of industrial collaborative control with a device-edge collaborative architecture. Based on a 5G industrial chip platform, it constructs an industrial AI inference and control system where multiple industrial terminals operate in wireless coordination with base stations and edge computing platforms on fixed cycles. Lightweight AI models are deployed on the terminal side to handle real-time tasks such as image recognition and anomaly detection. Complex computations like multi-object tracking and semantic segmentation are offloaded on-demand to the MEC platform. The system supports remote model loading and differential updates. Terminals can adaptively switch inference modes based on network status and task complexity. 5G wireless communication replaces traditional wired connections, meeting the industrial control requirements for ultra-reliability, ultra-low latency, and deterministic communication. This solution achieves a dual enhancement in industrial collaborative control efficiency and flexibility. Task offloading reduces the computational burden and power consumption on terminals, promoting the lightweight design of industrial terminals and making them suitable for resource-constrained environments. It endows terminals with autonomous perception and decision-making capabilities, providing intelligent support for multi-robot collaboration. It optimizes bandwidth utilization efficiency, ensuring the transmission of critical data in high-concurrency scenarios. It improves operational flexibility, allowing AI capabilities to evolve with business needs without frequent hardware replacement, thereby aiding the flexible production upgrade of smart factories.

### ❸ Mechanical Guide Dogs

device
edge
network
cloud

Guiding visually impaired individuals requires a solution that is safe, reliable, and inclusive. Traditional guide dogs have long training cycles and high costs, making it difficult to scale and meet the vast demand. Mechanical guide dogs must simultaneously satisfy stringent latency requirements for obstacle avoidance and emergency stopping (crucial for personal safety), as well as constraints on computing power, energy consumption, and cost for long-term, portable use. Traditional device-side intelligence solutions relying on high-performance chips cannot balance reliability and affordability.

Mobile AI employs a "device-edge-network-cloud" collaborative architecture to resolve this core challenge, achieving demand balance through hierarchical deployment of intelligent capabilities. The terminal side retains only local closed-loop control functions such as emergency braking and basic obstacle avoidance, ensuring rapid response in critical risk scenarios. High-computation tasks such as complex environmental understanding, multimodal perception fusion, and multi-turn voice interaction are offloaded to the edge and cloud via the mobile network. This avoids prolonged high-load operation on the terminal, optimizing energy efficiency while maintaining safety baselines. This achieves a multiple value closed loop encompassing safety, inclusivity, and scalability. Terminals do not require high-performance GPUs, significantly reducing device manufacturing costs and power consumption, markedly extending battery life, and enhancing practicality. The combination of device-side local closed-loop control and network-side computing power assurance gives key decisions like obstacle avoidance and emergency stopping predictable latency, fully meeting the demands of safety-sensitive scenarios. Leveraging the wide coverage of mobile networks enables consistent intelligent experiences across urban areas. This elevates guide services from a single-device capability to an operable, scalable network-based service, facilitating the implementation of inclusive AI in scenarios serving special needs groups.

## 3.2 Mobile AI Smartphones, Wearable Devices, and Agents

To establish a clearer and more intuitive understanding of intelligent terminals, foster a unified industry consensus on the capability evolution of AI terminals, and guide the intelligent evolution of terminal systems, the intelligence level of terminals is categorized into five grades—Level 1 (L1) to Level 5 (L5)—based on the complexity of tasks they can execute within their functional scope and their degree of automation. A higher level indicates a greater intelligence level of the terminal.

**L5** Fully Autonomous Level)

Possesses capabilities surpassing human experts, achieving comprehensive autonomy.

**L4** Expert Level

Capable of providing customized services at the level of a human expert.

**L3** Assistive Level

The terminal can comprehensively understand user commands and intent, possesses the ability to proactively clarify intent, and has strong reasoning capabilities enabling automatic task decomposition and orchestration; possesses the ability to dynamically select and automatically call tools, content generation ability in at least one modality (e.g., text, audio, or image); and has both short-term and long-term memory capabilities.

**L2** Tool-Assisted Level

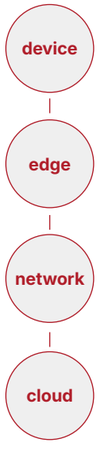
The terminal understands user commands and simple intent, possesses basic reasoning capabilities; has the ability to call preset tools to complete single-step or clearly defined multi-step execution tasks; possesses content generation ability in at least one modality (e.g., text, audio, or image); and has short-term memory within a single session.

**L1** Reactive Level

The terminal understands single, direct commands, invokes predetermined tools based on the command, and completes single-step tasks.

Currently, L3 AI agents capable of autonomous task closure are accelerating their development. It is projected that by 2030, terminal systems are expected to achieve L4 capabilities.

### ❶ Smartphone Terminal Agents





Terminal agents have become the core entry point for human-computer interaction, with their application scenarios continuously expanding. They exhibit a trend of collaboration between system-level agents and domain-specific agents, driving the transformation of apps towards "intent-centralization" and the AI-driven re-architecture of OS frameworks. This represents a leap from "command response" to "intent understanding." However, they face common industry challenges such as interaction logic transformation, protocol standardization, and cross-entity collaboration.

Mobile AI, with its core architecture centered on “device-edge-network-cloud” holistic collaboration, constructs an efficient operational and ecosystem synergy system for terminal agents. The system-level agent, acting as the "intelligent core," deeply integrates terminal hardware and operating system capabilities. Using purely on-device or device-cloud hybrid AI models, it accurately parses user multimodal intent from voice, text, gestures, etc. Combined with user profiles and scenario characteristics, it decomposes complex tasks and dynamically assigns them to vertical domain agents or application agents for execution based on task priority. Leveraging model quantization, pruning, and other lightweighting techniques, it adapts large models to the limited computing power of terminals, enabling local real-time response for high-frequency basic tasks. Complex tasks are offloaded to the edge or cloud for processing via the mobile network. In weak-network environments, it automatically switches to on-device inference to ensure service continuity. Simultaneously, it promotes the re-architecture of the OS framework around AI, placing large models and agents at the system's center, encapsulating atomic service capabilities, and supporting efficient invocation of multiple agents through unified interfaces. It facilitates the evolution of agent interaction protocols towards unification and standardization, leveraging protocols like A2A and MCP to establish channels for device interconnectivity, data sharing, and model collaboration, breaking down collaboration barriers across vendors and devices. The system agent also embeds secure and trustworthy mechanisms, employing data encryption and hierarchical permission control to ensure privacy and security during intent execution and data interaction. This solution drives a fundamental transformation in human-computer interaction paradigms, achieving a shift from "command-centric" to "intent-centric," delivering an integrated and seamless intelligent service experience. It propels terminal agents towards greater autonomy, personalization, and collaboration, reshapes the business logic of "hardware + intelligent services," contributes to the prosperity of an open and collaborative agent ecosystem, and accelerates the widespread adoption of Mobile AI on the consumer side.

### ❷ Cloud Agent for Inclusive Intelligent Terminals

**native terminal agent framework**

**+**

**scenario-specific ecosystem services**

The demand for inclusive AI is driving the intelligent upgrade of mid-range and entry-level terminals. There is an urgent need among middle- and low-income groups for cost-effective, easy-to-use, and reliable intelligent technology. Democratizing AI can help bridge the digital divide. However, enabling a vast number of mid- and low-end terminals to possess stable and sustainably evolving agent capabilities, without significantly increasing hardware costs, has become a core industry challenge.

Mobile AI addresses this challenge by constructing a cloud agent architecture. This solution deeply integrates agent capabilities at the operating system layer, forming a "native terminal agent framework + scenario-specific ecosystem services" architecture. Lightweight models are deployed on the terminal side to handle common intent recognition, simple reasoning, and execution tasks, ensuring local response for high-frequency demands. The cloud side, coordinated via the mobile network, hosts capabilities for complex planning, large model inference, and knowledge enhancement. Leveraging device-edge-cloud collaboration and unified interface orchestration, it enables cross-application, cross-service agent cooperation. Simultaneously, it integrates multimodal interaction technologies to improve scenario understanding and decision-making efficiency. This solution effectively promotes the widespread adoption of inclusive AI, empowering mid- and low-end terminals with intelligent service capabilities without requiring additional hardware costs. Completing high-frequency tasks on the device side reduces dependency on the cloud, allowing stable operation even in weak or no-network environments, thereby enhancing response speed and experience consistency. Cross-application collaboration breaks down experience fragmentation, driving agents to evolve from passive response to proactive and contextual services. This aids in narrowing the digital divide and enriching the supply of intelligent services within the scaled terminal ecosystem.

### 3 Light Mobile AI Agent Terminals with Global eSIM Connectivity

light device
heavy cloud


The global large-scale deployment of lightweight agent terminals faces dual demands of cost-effectiveness and service stability. With the advancement of large language models and multimodal AI technologies, traditional functional hardware is evolving into lightweight agent terminals capable of understanding and reasoning. However, a core industry challenge is ensuring the stable and continuous delivery of intelligent capabilities during their global deployment, without significantly increasing hardware costs.

Mobile AI tackles this problem by combining a "light device + heavy cloud" architecture with global eSIM connectivity. The core concept is to centrally deploy high-computing-power tasks (such as multimodal interaction, large model inference, and knowledge base retrieval) on the cloud side, delivering them to terminals on-demand, securely, and with low latency via the mobile network. On the terminal side, eSIM capability compliant with GSMA standards is integrated into the cellular communication module, enabling global automatic network connection and service access right out of the box. This eliminates the need for high-performance GPUs or NPUs, controlling hardware costs through compute offloading. Leveraging eSIM's intelligent multi-operator switching function, terminals can automatically match the optimal network path and the nearest inference node. This solution achieves a unification of low cost and globally stable service, significantly reducing terminal hardware and power consumption costs, allowing even affordable consumer electronics to deliver high-level intelligent experiences. The global connectivity capability of eSIM ensures service continuity for devices across regions and multiple network environments, meeting the low-latency and high-stability requirements of AI-powered real-time interaction. This propels the globalization and democratization of Mobile AI agent terminals.

### 4 Multi-Device Collaboration and Capability Extension

core terminal
edge peripherals
cloud support

Lightweight terminals face the challenge of balancing resource constraints with intelligent functionality demands. The rise of Generative AI is driving the evolution of terminals into agents, with wearables like children's smartwatches becoming portable interaction hubs. Future scenarios will encompass diverse form factors such as smart glasses and earbuds. However, a single device, limited by its size, battery life, and computing power, struggles to handle complex AI inference tasks. It is necessary to achieve device-cloud collaboration and shared computing resources within these constraints.

Mobile AI builds a distributed agent system based on a "core terminal + edge peripherals + cloud support" architecture. Using a smartwatch as the on-device computing and connectivity hub, lightweight multimodal models are deployed. These models aggregate visual, audio, and biometric data from various peripherals via short-range communication. The cloud side leverages AI capabilities to provide complex services such as in-depth companionship and learning assistance. Dynamic scheduling optimizes the energy efficiency of the "hub and peripherals," while device-cloud collaboration shares the computational load. The system also supports the local retention of agent capabilities, ensuring service continuity across multiple scenarios. This solution effectively overcomes the physical limitations of lightweight terminals. The distributed architecture extends the battery life of the entire ecosystem, and multimodal data collaboration enables natural and precise interactive experiences. In scenarios with poor or no network connectivity, the core terminal can still drive basic services, maintaining a closed loop for privacy and security. This promotes the evolution of wearable devices from single-function tools into a collaborative intelligent system, laying the groundwork for round-the-clock proactive services and facilitating the implementation of Mobile AI in wearable scenarios.

### 5 Summary

Mobile AI is centered on "device-edge-network-cloud" collaboration from an overall architectural perspective. Through the hierarchical deployment and dynamic scheduling of terminal and agent capabilities, it enables terminals of different forms and performance levels to obtain a stable and sustainable intelligent experience within their respective constraints. The terminal side assumes core capabilities such as secure closed-loop operations and immediate response. The network side ensures stable service delivery through wide-coverage connectivity and QoS guarantees. The edge and cloud sides centrally host complex inference and knowledge enhancement functions. Together, these three layers form the foundational system supporting the operation of agents and terminals.

Examining typical application practices, while the focus of Mobile AI varies for different types of agents, the common requirements are highly consistent. Security-sensitive scenarios primarily ensure the reliability of critical decisions. Industrial collaborative control scenarios focus on multi-terminal concurrency and deterministic latency. Resource-constrained terminals rely on device-cloud collaboration to balance experience and cost. Large-scale deployment scenarios achieve rapid coverage and consistent delivery of intelligent capabilities through system-level integration.

## 4. Mobile AI applications

Mobile AI leverages its core characteristics of ubiquitous intelligence, ubiquitous connectivity, and distributed intelligent collaboration to break through device and spatial limitations. It drives the deep penetration of intelligent applications across both the consumer (ToC) and enterprise (ToB) sectors. In the ToC domain, by expanding the dimensions of human-computer interaction and revolutionizing immersive experience paradigms, it integrates intelligent capabilities such as AR/VR, real-time speech translation, and customized lifestyle services into daily scenarios. This precisely matches user needs, significantly enhancing digital consumption experiences and quality of life. In the ToB domain, utilizing its technical advantages of high reliability and low latency, it empowers vertical industries like smart manufacturing, transportation & energy, and healthcare. Through optimizing production processes, enabling precise scheduling, and fostering collaborative innovation, it substantially boosts production efficiency and industrial resilience. This chapter will focus on key industries in both the ToC and ToB sectors, exploring how Mobile AI, through its mobile, pervasive intelligent capabilities, addresses pain points in traditional industries and drives industrial digital and intelligent transformation.

### 4.1 User Intelligent Consumption Experience

Current device-based intelligent experiences face distinct pain points: high-end intelligent services are constrained by both device capabilities and spatial reach, making universal access difficult. Additionally, the dimensions of human-computer interaction remain narrow, and scenario boundaries are rigid. Mobile AI, relying on its technologies of ubiquitous connectivity and deep distributed intelligent collaboration, effectively overcomes these bottlenecks. It not only expands the dimensions of interaction and scenario boundaries but also enables high-end intelligent experiences to transcend device and spatial limitations, achieving universal access and pervasive availability. This ultimately ushers in a new intelligent phase characterized by more immersive perception, more precise services, and more efficient living.

Digital Consumption Upgrade and Immersive Experience Innovation. Mobile AI, built upon real-time rendering, multimodal fusion technologies, and a technical architecture featuring efficient collaboration between cloud-based large models and terminals, also optimizes the interaction logic of intelligent services. It drives immersive applications like AR/VR fully into everyday life, constructing a new ecosystem of virtual-physical integrated social entertainment and digital shopping. It enables the practical implementation of intelligent capabilities like real-time speech translation and contextual image recognition and narration, breaking down language communication and cognitive barriers. This allows complex intelligent services to reach users in more natural and convenient ways, reshaping the paradigm of digital consumption experiences.

Enhancement of Social Efficiency and Personal Life Quality. Mobile AI leverages its extensive connectivity, combined with context awareness and intent understanding technologies, to link with IoT devices across domains and build solutions. On one hand, it precisely discerns user needs and provides customized services, such as meeting summarization and multilingual transcription in mobile office scenarios, and dynamic route optimization in smart mobility. On the other hand, it automates the handling of cross-scenario tasks, covering everything from smart home adjustments to workflow optimization. This ultimately helps users break free from repetitive tasks and focus on creative activities, effectively enhancing both societal operational efficiency and personal life quality.

Development of Diverse Terminals and Proactive Intelligent Services. Mobile AI employs computational offloading technologies to break the "large-screen centric" limitation. Relying on multimodal sensors and autonomous decision-making systems, it bridges the gap for deploying digital intelligence into physical scenarios. Simultaneously, using device-edge-cloud closed loops and cross-terminal collaboration technologies, it undertakes the core mission of seamlessly connecting personal AI devices (e.g., phones, smart glasses) with physical AI carriers (e.g., smart cars, humanoid robots). This successfully fosters a user-centric digital service layer, creating a coordinated device ecosystem centered on the user. It propels intelligent agents to evolve from passive response to proactive perception and demand prediction, realizing preemptive intelligent service experiences that comprehensively adapt to all-scenario intelligent demands.

## 4.2 Smart Manufacturing

Smart manufacturing has an urgent and stringent demand for mobile intelligent applications. Production sites have a critical need for intelligent applications such as equipment predictive maintenance, AR-assisted assembly, mobile inspection, and flexible scheduling. However, the traditional cloud-based AI model faces pressures from bandwidth and transmission costs. Millisecond-level delays can pose safety risks. Coupled with factory electromagnetic interference and data privacy protection requirements, relying solely on public cloud services struggles to meet these scenario demands.

Mobile AI adopts a "device-side inference + edge collaboration" core architecture. Terminals, leveraging built-in NPU computing power, can complete defect identification and anomaly alerts within milliseconds without relying on the cloud. Sensitive data undergoes cleansing and inference locally or at the edge MEC, with only result data being reported, firmly establishing privacy and security safeguards. For complex workshop environments with weak network coverage, terminals support basic recognition and recording functions, synchronizing data once the network is restored. The architecture is also adapted for industrial electromagnetic environments, enhancing technical stability and anti-interference capabilities to precisely match production scenario needs. This solution effectively meets the real-time and security requirements of smart manufacturing, reduces bandwidth consumption and data leakage risks, and improves usability in scenarios with poor network connectivity. It drives the evolution of production processes from "assisted operations" towards "unmanned operations," facilitates collaboration between autonomous mobile robots and intelligent agent terminals, lays the groundwork for "lights-out factories" and adaptive flexible manufacturing, and accelerates the digital and intelligent transformation process of the manufacturing industry.



## 4.3 Urban Governance

Urban governance is undergoing a transformation towards "smart" systems, with extensive application scenarios for mobile solutions. From grid-based mobile law enforcement and dynamic environmental monitoring to emergency response, various mobile terminals have become crucial supports for governance. However, the full backhaul of massive unstructured video data easily causes network congestion and storage pressure. Furthermore, data collection in public spaces involves citizen privacy, making real-time anonymization at the collection point a core challenge.

Mobile AI addresses these governance pain points through "computing capability prepositioning," extending intelligent processing power to the very edge of governance operations. Terminals equipped with AI recognition functions capture and upload evidence only upon detecting anomalous events, drastically reducing invalid data transmission. At the data collection source, on-device AI enables automatic blurring of private information, implementing preemptive privacy risk control. Inspection terminals transform into intelligent agents capable of automatically generating and distributing work orders, streamlining manual entry processes. Simultaneously, leveraging edge-side computing support ensures fast local data processing and response. This solution effectively reduces network and storage costs, fortifies privacy and security defenses, and significantly improves the efficiency of urban governance incident handling. It promotes the upgrade of urban governance from "sensing and discovery" to "cognitive prediction," aiding the extension from passive post-event response to proactive pre-event prevention, providing core technological support for building a refined and efficient modern urban governance system.

## 4.4 Intelligent Transportation

Intelligent transportation must balance safety and efficiency, with applications covering vehicles, roads, people, and events. Demands for intelligent connected vehicles, roadside sensing collaboration, and logistics tracking are increasingly urgent. However, high-speed vehicle movement can affect signal stability; weak-network environments like tunnels and mountainous areas pose safety risks; and the cost of fully uploading terabytes of vehicle-generated data to the cloud is prohibitive and impractical, forming a major bottleneck for industry upgrading.

Mobile AI tackles the pain points of transportation scenarios with a Vehicle-Road-Cloud integrated architecture, constructing a comprehensive safety and efficiency assurance system. Vehicles rely on onboard edge intelligence, utilizing powerful local computing to handle perception and decision-making for driving, ensuring core safety capabilities like emergency braking and obstacle avoidance even in no-network or weak-network conditions. User smartphones serve as the Mobile AI entry point, enabling seamless transfer of navigation and entertainment information to the vehicle infotainment system, and can also act as digital car keys for keyless entry. Utilizing C-V2X (Cellular Vehicle-to-Everything) technology, vehicles and roadside units engage in short-range communication and edge computing, achieving beyond-visual-range perception and effectively compensating for the limitations of standalone vehicle intelligence. This solution significantly enhances the safety and traffic efficiency of intelligent transportation, laying the foundation for autonomous driving development. Safety assurance capabilities in weak-network environments are greatly strengthened, all-scenario interconnected experiences are continuously expanded, and individual vehicle perception blind spots are effectively mitigated. It promotes the evolution of autonomous driving from L2+ towards L3/L4 levels, helping vehicles become independent intelligent agents that optimize traffic flow and alleviate congestion through swarm collaboration, giving rise to new Mobility-as-a-Service business models like "Robotaxi."

## 4.5 Smart Healthcare

Smart healthcare focuses on out-of-hospital and primary care scenarios, addressing the pain points of uneven resource distribution and inefficient emergency response. Core needs encompass 5G+AI-enabled emergency services, mobile-assisted diagnosis, and personalized health management. Challenges include the large size of medical imaging files, network fluctuations in emergency scenarios affecting remote consultation, a lack of expert-level diagnostic capabilities among primary care doctors, and stringent security requirements for medical data transmission.

Mobile AI builds a comprehensive healthcare intelligence system through terminal empowerment, data optimization, and proactive intervention. AI-assisted diagnosis terminals integrate algorithms

to help primary care doctors capture optimal examination views in real-time and automatically measure key indicators, improving diagnostic accuracy. In pre-hospital emergency care, edge AI devices preprocess and compress vital sign data, prioritizing the transmission of critical alarm information, while onboard AI models provide immediate first-aid recommendations. Wearable devices, relying on on-device models, analyze data trends to provide early health risk warnings, simultaneously ensuring medical data privacy through secure transmission technologies. This solution effectively improves healthcare accessibility and emergency response efficiency, promoting a transformation in health management models. It significantly alleviates the uneven distribution of medical resources, reduces misdiagnosis rates at the primary level, and creates a seamless "from ambulance to hospital admission" emergency life channel, shifting the focus from "treating disease" to "preventing disease." It provides technological support for addressing medical resource shortages in an aging society, with the potential to foster AI-powered family doctors in the future, further broadening the coverage and service depth of smart healthcare.

## 4.6 Energy & Power

The energy sector faces complex and challenging inspection and maintenance scenarios. Core requirements include automated inspection via drones/robots, refined quality inspection of cables and equipment, and smart meter reading with energy efficiency management. However, public network coverage for field facilities is weak, making real-time transmission of high-definition video difficult; traditional manual inspection is inefficient and prone to fatigue; and high-voltage electromagnetic environments impose stringent demands on communication anti-interference capabilities.

Mobile AI centers on on-device intelligence and scenario adaptation to build dedicated solutions for the energy industry. Portable AI-powered quality inspection all-in-one machines, combining 5G-A and vision algorithms, can perform 360-degree cable imaging and on-device inference in weak-network or offline environments, accurately identifying minor defects. Inspection drones equipped with edge computing modules analyze images in real-time, saving evidence and triggering alarms only upon detecting anomalies. Explosion-proof robots replace humans in high-risk areas, using on-device AI to read instrument gauges and detect thermal imaging anomalies, while featuring reinforced anti-interference designs to adapt to high-voltage environments. This solution greatly enhances the efficiency and safety of energy sector inspections and maintenance. Inspection efficiency is multiplied compared to traditional manual methods, significantly reducing the risk of missed defects and personnel safety hazards, while being adaptable to field weak-network and high-voltage electromagnetic conditions. It promotes the evolution of energy systems towards "self-healing grids." Future deployment of miniature Mobile AI sensors will enable autonomous fault localization and isolation, continuously enhancing the resilience and intelligence level of energy systems.

## 4.7 Summary

From an overall architectural perspective, Mobile AI relies on the core support of holistic "device-edge-network-cloud" collaboration. Leveraging the synergistic interplay of its "Three-Layer, Four-Dimension" framework, it transforms the network and computing capabilities of the foundation layer and the terminal execution capabilities of the execution layer into scenario-specific value at the application layer. The device side focuses on real-time perception and local response; the edge side handles regional collaboration and low-latency processing; the cloud side undertakes complex computation and global optimization; and the network side provides ubiquitous connectivity and precise scheduling. Together, they constitute the complete technical support system for application implementation.

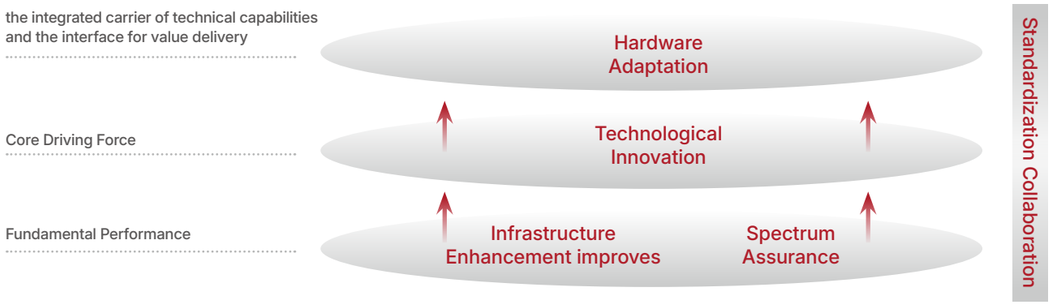
Examining typical application practices, the emphasis of Mobile AI requirements varies across different industries, yet the core demands are highly unified. Smart manufacturing focuses on low latency and data security. Urban governance emphasizes efficient data processing and privacy protection. Intelligent transportation relies on vehicle-road-cloud collaboration and safety assurance. Smart healthcare highlights resource allocation to grassroots levels and emergency efficiency. Energy & Power stresses adaptation to harsh environments and operational safety. The commonality is that all leverage technology to address scenario-specific pain points, achieving efficiency gains, cost optimization, and service upgrades, thereby driving the transformation of all industries from traditional models to digital and intelligent paradigms.



# 03 Towards Mobile AI

**To achieve the large-scale deployment and value realization of Mobile AI, it is essential to establish a systematic, multi-level support system. Based on industrial development needs and technological evolution trends, this chapter focuses on five critical pathways: infrastructure evolution, core technological breakthroughs, terminal ecosystem synergy, spectrum resource assurance, and collaborative standard development. It systematically outlines the comprehensive framework and potential solutions for transitioning Mobile AI from a technical concept to industrial application. Specifically, these pathways include: Infrastructure Enhancement, Spectrum Assurance, Technological Innovation, Hardware Evolution, and Standardization Formulation. These five interrelated and co-evolving dimensions collectively constitute a complete system that supports the transformation of Mobile AI technological capabilities into industrial practice.**

Infrastructure Enhancement improves the fundamental performance of networks and computing power, providing the physical layer for Mobile AI characterized by high reliability, low latency, and ubiquitous connectivity. Spectrum Assurance, through the scientific planning and dynamic management of spectrum resources, provides the essential wireless transmission resources for diverse AI services. Together, they constitute the resource and capability foundation for Mobile AI operation. Technological Innovation acts as the core driving force, transforming foundational resources into systemic capabilities. Centered on the bidirectional empowerment of AI and networks, and through the design of an AI-native network architecture combined with end-to-end secure and trustworthy control, it becomes the key enabling layer bridging infrastructure and upper-layer applications. Hardware Adaptation serves as the integrated carrier of technical capabilities and the interface for value delivery. Through hardware upgrades, industrial collaboration, and ecosystem co-creation, it converts network capabilities, computing power, and intelligence into user-perceivable services, directly influencing the experience quality and application scale of Mobile AI. Standardization Collaboration runs through all the aforementioned aspects. By establishing unified technical specifications, interface protocols, and evaluation systems, it promotes interoperability and efficient collaboration across all layers of the industrial chain, reduces integration complexity, and accelerates the maturation and adoption of holistic solutions.




## 1. Infrastructure Enhancement

Infrastructure Enhancement serves as a critical cornerstone for the large-scale deployment of Mobile AI. It aims to construct an underlying guarantee system characterized by "adaptive connectivity, sufficient computing power, and efficient collaboration" through the upgrading of fundamental communication network capabilities and the optimized allocation of computing resources. This precisely matches the differentiated demands of Mobile AI's diverse scenarios, overcoming the constraints of traditional infrastructure in terms of performance, capacity, and flexibility.

Enhancing fundamental communication network capabilities to dynamically adapt to diverse scenario demands. Communication networks must evolve from "generic connectivity" towards "scenario-customized connectivity," upgrading around four core dimensions: high bandwidth, low latency, massive connectivity, and high reliability, while establishing dynamic adaptation mechanisms. Transmission capacity is increased through multi-band coordination and carrier aggregation to support demands from scenarios like AI large model data transfer and AR/VR. The network protocol stack is optimized, and edge nodes are introduced to reduce end-to-end latency for scenarios such as intelligent connected vehicles and telemedicine. Innovative technologies enhance connection density to meet the demand for concurrent access by millions of terminals. Reliability for critical industry scenarios is ensured through redundant transmission and link self-healing. A business-aware and resource scheduling linkage system is built to dynamically match the real-time needs and priorities of different services.

Computing power enhancement to adapt to all-scenario computing demands. Addressing the differentiated computing power requirements of Mobile AI at the device, edge, and cloud sides, a layered, collaborative "device-edge-cloud" computing power supply system is constructed. The cloud side expands high-performance computing clusters to support large-scale training. The edge side increases nodes to enhance low-latency response capabilities. The device side strengthens the execution of local, lightweight models through dedicated AI chips. Heterogeneous computing facilities are integrated to form a unified resource pool, with elastic scaling space reserved to accommodate computing power growth. Intelligent energy consumption management, redundant backup, and computing power monitoring mechanisms are established to ensure stable supply. A diversified "general-purpose + specialized" computing power system is built to adapt to different scenario needs. Intelligent scheduling algorithms enable dynamic allocation of large models, improving computing resource utilization and service response speed.

## 2. Spectrum Assurance

Spectrum is a foundational resource for carrying Mobile AI service data and a key prerequisite for ensuring ubiquitous intelligent experiences. To meet the extreme demands for network capacity and determinism in scenarios such as AI large model interaction and multimodal data transmission, spectrum strategy must follow a dual evolutionary path parallelly pursuing "planning of new frequency bands" and "optimization of existing spectrum."

**prospectively plan new mid- and high-frequency band resources**

Promote mid-frequency bands, including 6 GHz, to become globally unified new primary bands, providing contiguous large bandwidth to primarily guarantee the high uplink demands of services like distributed AI inference and high-definition machine vision. Simultaneously, actively explore the deployment of high-frequency bands like millimeter wave in hotspot areas, leveraging their ultra-large capacity and ultra-low latency characteristics to empower cutting-edge applications such as immersive XR and holographic communication.

**continuously tap into the potential of existing spectrum resources**

By introducing AI-driven dynamic spectrum management, intelligent spectrum aggregation, and other solutions, achieve flexible scheduling and efficient sharing of cross-band, cross-radio access technology (RAT) resources. This addresses the challenge of massive concurrent access by AI terminals in scenarios like stadiums and smart factories. Combined with advanced technologies such as massive MIMO, spectrum reuse efficiency can be significantly enhanced, thereby maximizing overall network capacity and user experience under conditions of limited total spectrum resources.

# 3. Technological Innovation

Technological innovation is the core driving force for realizing Mobile AI. It transforms foundational resources into systemic capabilities, providing efficient support for the implementation of Mobile AI applications. The Mobile AI technological innovation system is built upon an AI-native network architecture. It develops the bidirectional technological systems of AI for Network and Network for AI, builds high-efficiency terminal carriers via technologies such as terminal heterogeneous intelligent computing and multi-agent collaboration, and expands implementation scenarios through application innovations including multimodal interaction and scenario-based solutions, safeguarded by an end-to-end security technology framework, achieving the native symbiosis and bidirectional empowerment of networks and AI.

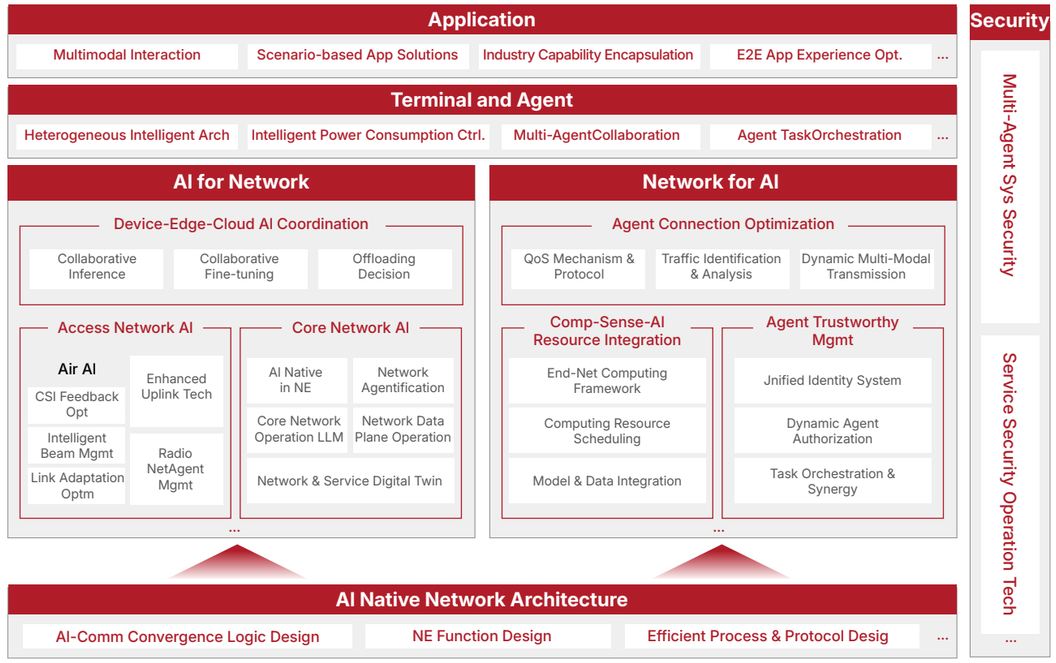



Figure 5: Mobile AI Innovation Technology Map

Since it covers a wide range of technical fields, this chapter will focus on discussing the core content related to the AI native network architecture, AI for Network, Network for AI and security.

## 3.1 AI Native Network Architecture

The AI-native network architecture for Mobile AI serves as the cornerstone for achieving bidirectional empowerment between networks and AI. Its core objectives are to support native AI functions, enhance network performance and efficiency, and provide ubiquitous, efficient communication connectivity along with new types of services (such as computing, data, and models) for business applications.

In the aspect of AI for Network, AI is integrated as an inherent gene from the initial architectural design, deeply fused into network function design and interface interaction definitions. Its core characteristics include three main features:

**Hierarchical Centralized Control:** Business-level centralized control and intent comprehension are implemented in the core network, while regional-level real-time control is performed in the radio access network, enabling precise and flexible multi-level network management and control.

**Functional Categorization and Aggregation:** Network functions are partitioned according to dimensions such as data and computing. They respectively undertake tasks like all-element data production and consumption, and integrated control and execution of computing-network convergence, collaboratively providing high-quality data and computing power support for AI models.

**Flexible On-Demand Deployment:** Distributed execution functions are dynamically deployed based on business scenario demands, enabling efficient and dynamic resource allocation and improving overall resource utilization.

In the aspect of Network for AI, the AI-native network architecture will serve as a unified platform integrating communication, sensing, computing, and intelligence to offer intelligent, integrated services externally. Fundamental network capabilities are encapsulated into standardized services and exposed through unified interfaces. AI applications can flexibly invoke the required services via the centralized control plane. The network then flexibly deploys service instances to the optimal nodes on-demand based on business performance requirements. The resulting model of "service exposure - capability invocation - proximity execution" can effectively support the stringent low-latency and high-computing-power demands of emerging applications like AI Agents and embodied intelligence.

## 3.2 AI for Network Key Technologies

AI for Network aims to deeply integrate AI into the end-to-end network system, making it a key production factor for networks, and to build an endogenously intelligent network capable of closed-loop "Perception-Cognition-Decision-Execution" capability across the entire chain. Its technological innovations span all domains of the end-to-end system, including the air interface, core network, security, and cross-domain collaboration.

### 1 Access Network AI

As Mobile AI deeply penetrates from the cloud to the edge and terminal sides, future communication networks are facing an unprecedented paradigm shift. Innovation in the air interface is urgently needed to reshape the connectivity performance for Mobile AI, with key paths primarily reflected in two major dimensions: Air Interface AI and Enhanced Uplink New Technologies.

**- Air Interface AI:**

By introducing deep neural networks, the network can perform adaptive learning and optimization for complex channel environments, significantly reducing air interface overhead and improving spectrum efficiency. Its core technical capabilities include:

**Channel State Information (CSI) Feedback Optimization**

AI algorithms can intelligently predict and compress channel state information, greatly reducing feedback overhead and improving spectrum utilization efficiency.

**Intelligent Beam Management**

AI algorithms enable intelligent beam selection and management, quickly finding the optimal beam configuration in complex and varying channel environments, thereby improving connection quality and coverage.

**Link Adaptation Optimization**

AI algorithms can dynamically adjust parameters such as modulation and coding schemes and power control based on real-time channel conditions and service demands, optimizing link performance.

**- Enhanced Uplink New Technologies:**

The "device-side perception, cloud/edge-side reasoning" end-to-end closed loop of Mobile AI imposes stringent requirements on uplink bandwidth and latency. Enhanced Uplink New Technologies serve as the physical foundation for Mobile AI to achieve device-cloud collaboration. Their core technical capabilities include:

**Spectrum Reconfiguration**

Through flexible spectrum allocation strategies, more spectrum resources are allocated to the uplink, enhancing uplink capacity.

**Flexible Slot Allocation**

The uplink/downlink slot ratio is dynamically adjusted based on service demands, increasing the proportion of uplink slots in Mobile AI scenarios to meet the needs of uplink data-intensive transmission.

**Multi-Band Coordination**

Spectrum resources from multiple bands are aggregated through multi-band carrier aggregation technology, further increasing uplink bandwidth.

Through Enhanced Uplink New Technologies, networks can break the uplink bottleneck, ensuring that high-value data from the terminal side can be uploaded to cloud computing centers in a timely manner, meeting the latency requirements of Mobile AI's end-to-end closed loop.

**- Wireless Network Agent:**

The core objective is to achieve autonomous closed-loop within the wireless domain. Relying on local domain data and global policy input, it performs near-real-time analysis, inference, and decision-making, supporting the autonomous operation, maintenance, and optimization of wireless networks. Its core characteristics are reflected in two aspects:

Single-Domain Autonomy, through a management-control integrated intelligent engine, it achieves local network scenario identification, risk prediction and prevention, fault root cause analysis, and end-to-end automated operation and maintenance optimization.

Agent Autonomous Closed-Loop, for specific scenarios such as operation and maintenance, energy saving, and experience assurance, it builds end-to-end autonomous capabilities of "Perception - Analysis - Decision - Execution."

The core enablers of this technology are the Communication Large Language Model and the Radio access network Digital Twin System (RDTS). The Communication Large Language Model, pre-trained and fine-tuned on large-scale telecom corpora and domain knowledge, possesses capabilities such as intent translation, multi-objective collaborative optimization, and proactive trend analysis. The RDTS achieves real-time digital representation of the physical network by constructing multi-dimensional digital mappings of all network elements. It provides high-precision modeling, strategy simulation and verification, and full-chain traceability capabilities, offering accurate support for intelligent decision-making in the wireless domain and significantly enhancing network autonomy levels and decision-making credibility.

## 2 Core Network AI

The core network serves as the center for service policy control and network topology. To adapt to personalized service demands and dynamic resource optimization, the core network urgently requires intelligent upgrades. It needs to support AI Native in Network Elements and Network Agentification, while using data and twin technologies to solidify the foundation for this transformation.

**- AI Native in Network Elements:**

The policy decision logic of traditional core networks is based on static rules, which cannot fully satisfy users' personalized and diverse demands, nor achieve optimal real-time dynamic configuration of network resources. It is necessary to introduce AI model-driven inference and decision-making. By embedding AI natively into the functional logic and interface interactions of network elements, network adaptive capabilities in generalized scenarios can be realized. AI Native in Network Elements is reflected both in AI being deployable down to the perception and reasoning/decision-making of each individual network element (e.g., service feature analysis or policy adjustment) and scalable up to network-level resource scheduling and configuration (e.g., dynamic slicing or network composition). To ensure the high reliability of network operation, AI Native in Network Elements is also reflected in the autonomy of AI models operating within the network. Although offline pre-trained models acquire basic generalization capabilities, inevitable quality biases in training data samples (e.g., class imbalance) and feature shifts in the deployment environment (e.g., differences in user or network resource distribution) necessitate that AI models can perform periodic or real-time-feedback-based online learning updates using data samples from the actual network deployment environment. This continuously optimizes model performance and scenario adaptability, ensuring the inference efficiency and accuracy of models running in the network.

**- Network Agentification:**

As the central nervous system of the communication network, undertaking core functions such as service subscription management, policy execution, and network element coordination, the core network can evolve into a Network Agent based on its foundational capabilities. This enables intent understanding of user/service demands, complex task planning, tool invocation, and autonomous closed-loop execution. Compared to general-purpose AI Agents, core network agents have specific capability requirements, Context Management introduces an independent data plane and advanced Retrieval-Augmented Generation (RAG) technology to extract key information; Planning and Reasoning requires the capability for real-time, reliable long-horizon task planning, orchestration, and closed-loop iteration; Tool Use involves the modular reorganization of network element functions, improving invocation accuracy through constrained decoding; Protocol Interworking involves defining the NetMCP protocol to achieve unified abstraction and standardized invocation of network capabilities. Furthermore, multi-agent collaboration technologies need to be introduced to decompose complex tasks. Concurrently, the telecom industry is promoting the standardization of the A2A-T (Agent-to-Agent for Telecommunications) protocol to build a unified cross-domain, cross-vendor collaborative framework. This framework parses natural language business intents into network-wide collaborative tasks, driving the network towards intent-driven and ecosystem-collaborative intelligent transformation.

**- Core Network Operation Large Model:**

Through real-time analysis and processing of network operation data (such as traffic, signaling, etc.), large models can comprehensively perceive key information including service quality, network load status, and user behavior. Leveraging their powerful feature extraction and correlation analysis capabilities, they can break through traditional rule-based discriminative decision-making methods, enabling efficient network operation and the provision of high-quality, personalized network services. Compared to smaller-scale network AI models, large models have significant advantages in generalization and accuracy. On one hand, through a "foundation model + local fine-tuning" approach, they enable rapid migration and adaptation of applications across regions. A unified model framework greatly reduces the cost of repeated model development. On the other hand, leveraging the massive parameters and high knowledge content of the models, they can integrate multi-dimensional network information to achieve comprehensive optimization decisions, making network operations more refined and accurate.

**- Network Operation Data Plane:**

The data plane is the core data hub supporting intelligent collaboration in the core network, bearing the key responsibility of managing the full data lifecycle and releasing its value. To better improve network operational efficiency and maximize the value of network data, it is necessary to construct an integrated data plane architecture through real-time data collection and intelligent analysis technologies, deeply transforming network data into a core asset driving business closed-loop operation. The network operation data plane needs to address four major technical challenges:

**Efficient Data Collection:**
Facilitating on-demand subscription, deduplication, and coordination for multi-source data, integrating data from network elements, terminals, service platforms, and other full-scenario sources.

**Real-time Data Processing:**
Improving data quality through intelligent cleansing, real-time aggregation, dynamic anonymization, and other technologies.

**Efficient Storage and Retrieval:**
Supporting efficient access to massive structured and unstructured data through hierarchical/classified storage and multi-dimensional correlation technologies.

**Secure, Trustworthy, and Open Sharing:**
Enabling secure and compliant data sharing through permission control and standardized interfaces.

The data plane enables efficient supply of network data, reduces redundant data collection, solves data silo problems, and promotes the network's transition from "rule-driven" to "data-and-AI self-adaptive." This helps operators transform into comprehensive "network + computing + intelligence" service providers, solidifies the data foundation for Mobile AI's evolution from converged application to native symbiosis, and accelerates the high-quality development of the digital economy.

**- Network Digital Twin and Service Twin:**

Core network AI will rely on core business data such as service experience data, configuration data, and network status data to build network digital twin models. These models accurately map the physical network's topology, resource status, and service flow paths, achieving high-fidelity representation of the physical network in digital space. Based on different modeling data domains and virtual-physical linkage objectives, they can be categorized into Network Digital Twin and Service Twin.

- **Network Digital Twin:** By building an AI simulation system for dynamic networks, it deeply integrates multi-dimensional sub-models such as terminal behavior, device operation, and fault evolution. Combined with technologies like task parallel scheduling and statistical vector optimization, it improves simulation efficiency and scenario fidelity. Based on network available path solving algorithms, it can quickly locate faulty network elements and paths, promoting the upgrade of operation and maintenance from "passively perceiving service degradation" to "actively isolating faulty network elements," providing core support for network high stability and operational efficiency improvement.
- **Service Twin:** Using big data statistical mining and large model simulation prediction as its dual core foundations, it focuses on diverse business scenarios such as service experience analysis, high-speed rail scenario analysis, and media relay. It delves into key capability enhancement paths like identifying service experience bottlenecks, service configuration optimization, and tariff plan launch evaluation, exploring the construction of an iterative enhancement ecosystem for "precise network planning/construction and efficient service implementation." Its core encompasses the technical optimization of the spatiotemporal multi-dimensional business analysis foundation, research on core algorithms for the AI intelligent twin foundation, the implementation path for large-scale deployment capability, the construction scheme for multi-dimensional core business views and comprehensive analysis systems, and tackling technical challenges for the full-process closed-loop management of platform-triggered evaluation, data subscription and reporting, and intelligent prediction presentation. It provides solid theoretical support and mature technical solutions for business innovation iteration and steady revenue growth.

### ❸ Terminal–Edge–Cloud AI Collaboration

The core objective of Terminal–Edge–Cloud collaboration is to achieve the optimal balance between business demands and the real-time status of network and computing resources, thereby dynamically selecting the most suitable execution location among the terminal, edge, and cloud. Based on 5G-A/6G networks, terminals can collaborate with edge or cloud AI to achieve distributed collaborative computing for complex tasks, transforming terminals into key holistic intelligent hubs that integrate environmental perception, network collaborative orchestration, local processing, and result presentation. The core technologies for Terminal–Edge–Cloud collaboration include Collaborative Inference, Collaborative Fine-tuning, and Network-Assisted Offloading Decision.

- **Collaborative Inference:** Collaborative Inference includes inference offloading and split inference. Inference Offloading: The terminal typically performs preprocessing first, then uploads compressed features to the edge or cloud for the main inference task, reducing link load and improving overall efficiency; By selecting appropriate split points in the model structure, Split Inference enables Terminal–Edge collaborative execution, achieving a dynamic balance between latency and accuracy.
- **Collaborative Fine-tuning:** Collaborative Fine-tuning refers to scenarios involving complex training and fine-tuning. By performing lightweight updates on the terminal side and completing high-complexity training on the edge or cloud, it can not only reduce data leakage risks but also improve model iteration efficiency.
- **Network-Assisted Offloading Decision:** Efficient Terminal–Cloud–Edge collaboration requires comprehensive consideration of multi-dimensional factors such as queuing delay, link quality, terminal battery and thermal status, privacy domain constraints, and business QoE. These factors need to be continuously updated during task execution to ensure the inference offloading strategy remains optimal. In mobile scenarios, the system can migrate inference sessions to closer nodes based on terminal location and network conditions, or fall back to local inference when necessary, to maximize service continuity.

## 3.3 Network for AI Key Technologies

As new terminals and applications within communication networks, the large-scale proliferation of intelligent agents is driving the upgrade of communication protocols and the transformation of network architecture. There is an urgent need for networks to introduce new technological systems to meet the requirements of agent communication for high concurrency, high performance, low latency, and high security. The agent communication technology system encompasses multi-dimensional technical capabilities such as connection optimization, computational empowerment, and trustworthy management.

### ❶ Agent Connection Optimization

The purpose of Agent Connection Optimization technologies is to meet the demands of agent interconnections for new traffic patterns, new interconnection paradigms, and new networking methods. This includes designing new QoS Mechanisms and Protocol Design, Traffic Identification and Feature Analysis, and Dynamic Multi-Modal Traffic Transmission technologies to fulfill the transmission requirements of new information elements (e.g., AI models, tokens), new traffic directions (e.g., east-west traffic between agents), and new traffic characteristics (e.g., multi-modal bursts). New mechanisms such as registration/discovery and on-demand wide-area interconnection are designed to support task-driven point-to-point network connectivity. A flexible, task-driven multi-agent networking mechanism is constructed to enable swarm intelligence collaboration among heterogeneous agents.

### ❷ Computational Empowerment

Computational Empowerment technologies aim to empower intelligence by enabling networks to provide new data and computing services. This includes designing an End-Network Collaborative Computing Service Framework that supports the management, selection, and offloading of computing resources. It selects and offloads agent computing tasks to appropriate computing resources at the terminal, network, edge, or cloud sides, meeting the efficient and massive computing demands of agent communication. The Collaborative Scheduling of General Computing Resources technology supports the efficient and collaborative scheduling of computing resources under complex multi-dimensional constraints such as latency, throughput, resource utilization, and computational load. An end-to-end unified data service framework is constructed to support full lifecycle data management and multi-modal heterogeneous data transmission, empowering requirements such as agent AI model training and data sharing.

### ❷ Agent Trustworthy Management

Agent Trustworthy Management technologies use a Unified Identity Identification System to associate and manage agents of different forms belonging to the same user. This includes a unified identity mechanism for embodied intelligence/agent applications, an agent-user association mapping mechanism, and a Dynamic Authorization of Agent Permissions mechanism based on task granularity (rather than static subscription), to achieve trustworthy and secure agent communication.

## 3.4 Network AI Security Key Technologies

The security of Mobile AI encompasses two main aspects: first, how to ensure the security of the Mobile AI system itself, and second, how to utilize Mobile AI to serve security operations.

### ❶ Multi-Agent System Security Technology

The security protection of multi-agent systems differs fundamentally from that of traditional communication systems. As these systems are deployed across various domains, their complexity and attack surface continue to expand. Protecting against risks such as direct and indirect prompt injection, model poisoning, and vulnerabilities in agent AI tool invocation, which Mobile AI faces,

has become a core concern for the industry. When Mobile AI possesses agent characteristics, multi-agent collaboration enhances system flexibility and capability, but it also allows attackers to leverage the interaction relationships between agents to launch complex attacks. The attack surface extends significantly from a single agent to the level of system coordination. Multi-agent security solutions cover the entire lifecycle of agents, from creation to decommissioning, with its core spanning the four stages of design, development, deployment, and operation. Following the governance principles of "shift-left security" and "continuous assurance," security measures are moved forward to the design and development stages, and their security compliance is continuously monitored and evaluated throughout the lifecycle. Based on the OWASP guidelines for threat modeling in multi-agent systems, the MAESTRO (Multi-Agent Environment, Security, Threat Risk, and Outcome) framework can be leveraged to systematically analyze security threats, risks, and potential consequences within multi-agent collaborative communication environments. This enables targeted countermeasures, effectively enhancing the security and reliability of Mobile AI systems.

### ❷ Mobile AI Technology for Assisting Security Operations

Generative AI is being widely applied in the field of cybersecurity to enhance defensive capabilities and is gradually being embedded into various security products. In the domain of specialized large models, dedicated models for security defense and models used for attack testing have already emerged. Simultaneously, generative AI is evolving from a general-purpose auxiliary tool into specialized security agents and solutions, promoting the intelligence of cybersecurity. Generative AI significantly improves efficiency in Security Operations Centers (SOCs), where it is already used to automatically or semi-automatically complete tasks such as alert classification, threat intelligence analysis, incident response, and rule generation. AI-driven network security reasoning systems can automatically discover and fix vulnerabilities in open-source code. As the capabilities of models in the communications field improve, they can lower the cost and barrier to security testing. In the future, more frequent, proactive, and comprehensive security testing and patch management will be possible. For defenders, this will bring about a significant transformation in security operations and maintenance models, propelling cybersecurity into a new stage characterized by intelligence and high frequency.

## 4. Terminal Evolution

Terminal is the final carrier through which Mobile AI services reach users and industries. The evolution of its capabilities and the maturity of its ecosystem directly determine the boundaries of the ubiquitous AI experience and the scale of commercialization. The core objective of terminal support is to adapt to diverse scenario demands through hardware capability upgrades, build a standardized ecosystem through industrial collaboration, propel Mobile AI terminals from pilot exploration to large-scale commercial deployment, and provide solid support at the terminal level for the implementation of the entire industry chain.

**Hardware Upgrades to Adapt to Differentiated Demands Across All Mobile AI Scenarios**

The diversification of Mobile AI applications places comprehensive demands on terminal hardware that go beyond traditional communication requirements, necessitating evolution towards multimodal sensing, full-band reliable connectivity, and efficient local intelligent processing. Terminals need to integrate and efficiently process multimodal data such as voice and vision, balancing sensor accuracy, power consumption, and integration density to provide a unified sensing foundation for various AI applications. Existing and future mobile communication frequency bands require comprehensive optimization, strengthening 5G-A support while ensuring compatibility with new 6G bands. Ubiquitous coverage is built through the intelligent convergence of 5G with Wi-Fi and satellite communications. Research into AI model lightweighting and the integration of dedicated AI processing units should be deepened, promoting the migration of inference and decision-making capabilities to the terminal, forming an intelligent tiered system to support local real-time intelligence and reduce reliance on the cloud.

**Innovate terminal forms to unlock the service potential of Mobile AI across diverse scenarios**

To meet the development demands of Embodied AI and Inclusive AI, terminal forms are evolving towards diversification, collaboration and ubiquity, forging a full-scenario adaptive terminal system. We prioritize the innovation of embodied intelligent terminals, the R&D of user-centric terminal clustering and lightweight ubiquitous intelligent terminals to enrich the terminal product matrix. Meanwhile, we take multi-form terminals as carriers to foster brand-new service models, drive business innovation through the innovation of terminal forms, and activate the application value of Mobile AI in various fields.

**Full Industry Chain Collaboration to Facilitate Large-Scale Commercialization of Mobile AI Terminals**

Terminal manufacturers, chip companies, and operators need to establish deep collaboration mechanisms to address bottlenecks like fragmented technical standards and build an industrial landscape characterized by "unified standards, co-creation of features, and shared ecosystem prosperity." These three parties should jointly formulate Mobile AI terminal technical specifications, clarifying core metrics such as power consumption, latency, and connection reliability, and unifying interface protocols and adaptation standards to reduce collaboration costs across the industrial chain. Cross-industry innovation platforms should be established to share R&D outcomes and scenario requirements, forming a collaborative innovation chain of "chip-device-network-application." Operators leverage their network and scenario advantages, while terminal manufacturers and chip companies focus on hardware R&D, promoting the deep integration of software, hardware, and network capabilities, accelerating terminal technology maturation and commercial proliferation.

## 5. Standard Formulation

Standard unification is the institutional guarantee for breaking down Mobile AI industry collaboration barriers and promoting scaled development. By standardizing technical interfaces, unifying application specifications, and clarifying evaluation dimensions, it reduces industrial chain collaboration costs, accelerates scenario implementation, ensures consistent user experience, and provides solid support for the industry's healthy and orderly development.

**Promoting Globally Unified Technical Standards to Foster Industry Consensus**

Focusing on the collaborative needs of Mobile AI across layers, devices, and scenarios, efforts should be made to promote the establishment of a globally unified technical standard system. Key areas include standardizing the interconnection interface protocols between terminals, edge, network, and cloud, defining multimodal interaction frameworks and data transmission specifications, and providing a unified technical language for complex applications like AI agents, embodied intelligence, and intelligent connected vehicles. Through the concerted efforts of international standardization organizations and industry alliances, global technical consensus can be forged, collaboration obstacles caused by technical fragmentation can be eliminated, compatibility and adaptation between products from different vendors can be ensured, R&D and adaptation costs across the industrial chain can be reduced, and global industry collaboration efficiency and experience consistency can be enhanced.

**Perfecting Industry Application Standards to Accelerate Standardized Scenario Implementation**

Targeted at the differentiated needs of key industries such as energy, construction, manufacturing, healthcare, and transportation, industry-specific application standards should be formulated. Industry standards guide Mobile AI applications towards normalized, standardized development, reducing redundant R&D and trial-and-error costs, bridging the adaptation channel between technology and industry demands, propelling solutions from pilot validation to large-scale replication, and unleashing the value of industry digital and intelligent transformation.

**Building Quantifiable Experience Evaluation Standards to Align with Full-Chain Value**

Industrial chain resources should be integrated to construct a quantifiable, implementable Mobile AI experience evaluation system covering scenario modeling, metric definition, and capability adaptation. A unified, objective evaluation standard for multimodal interactive experience should be established, addressing issues like the low efficiency of subjective evaluation and insufficient quantitative basis. Clear mapping rules between AI application Key Quality Indicators (KQI) and Key Performance Indicators (KPI) such as network and terminal performance should be defined to achieve precise alignment between experience requirements and technical capabilities. Scenario-based test cases should be standardized, incorporating full-chain factors like network transmission, terminal performance, and algorithm efficiency into the evaluation scope to ensure evaluation results align with genuine user perception. Building upon the project foundations of existing standardization organizations, the AI-MOS (AI Mean Opinion Score) evaluation system should be continuously improved to provide an objective basis for industry optimization and upgrading.

# 04

# Recommendations and Call to Action



With the deep convergence of artificial intelligence and mobile communication technologies, Mobile AI has become a foundational engine for advancing the global digital economy and accelerating the digital-intelligent transformation. Two major trends are shaping its evolution.

First, the emergence of AI-Native Network Architecture has established the dual pathway of Network for AI and AI for Network , forming the early prototype of self-evolving intelligent systems.

Second, the rapid proliferation of AI-powered terminals - ranging from AI phones and AI glasses to AI vehicles, robots, and even AI toys - is shifting massive traffic towards the network edge, making edge intelligence an irreversible industry direction.

At the same time, the Internet of Agents is beginning to take shape, and Mobile AI-enabled applications are accelerating across vertical industries.

From a commercial perspective, operators are moving from traffic-centric operations to experience-driven monetization. The integrated "network + computing + data" service architecture enabled by Open Gateway is becoming mainstream. New business models – such as tiered billing, capability orchestration, and ecosystem expansion - are emerging, paving the way for scalable, repeatable commercial value loops.

Mobile AI now stands a **pivotal moment – advancing from technical feasibility to commercial viability.** Yet key challenges remain.

| Standards | Infrastructure | Commercialization |
|---|---|---|
| Lack of unified evaluation systems for AI products, inconsistent data formats, unstable agent protocols, and immature API specifications are resulting in fragmentation and rising integration costs. | Uplink bottlenecks, surging east-west traffic, and the stringent latency and elasticity demands of distributed AI impose significant pressure on current network architectures. | Value-sharing mechanisms remain underdeveloped. Many solutions do not adequately address industry pain points, and unclear application scenarios hinder the release of AI's full economy potential. |

To address these challenges and unlock Mobile AI's transformative value, we issue the following six industry-wide recommendations:

**Foster Strategic Consensus**

The growth of Mobile AI is both a technological inevitability and the product of coordinated policy and industry efforts. Policymakers should adopt open and inclusive approaches, optimize spectrum planning, activate data resources, strengthen digital infrastructure, and refine governance framework to ensure a stable and predictable environment. Operators should accelerate their evolution into TechCos, building core capabilities in computing-network orchestration, open platforms, and value distribution. Terminal manufacturers should deepen device-edge-network-cloud collaboration and cultivate robust agent ecosystem. AI companies must focus on real-world scenarios and deliver lightweight, accessible solutions. Industry users should contribute high-quality data, clarify scenario requirements, engage in standardization, and collectively drive industry-wide progress.

**Build a Globally Unified Standardization System**

Leverage the 6G standardization window to avoid fragmented development. International organizations such as GSMA, GTI, ITU, and 3GPP should jointly drive unified standards including:

- Establishing intelligence grading standards for AI terminals.
- Define labeling rules for AI-generated content.
- Formulate remote safety control requirements for AI vehicles and robots.

- Standardize AI model invocation methods, agent protocols, tool-calling APIs, and multimodal data formats.
- Create unified specifications for cloud-network-edge-terminal computing power collaboration and model streaming to adapt dynamically to network quality.
- Promote standardized APIs for exposing network, computing, and data capabilities to enable new business models and unlock new business values.

| | |
|---|---|
| **Reserve and Optimize Spectrum Resources:** | Conduct forward-looking spectrum planning with emphasis on uplink expansion. Accelerate the deployment of 6 GHz and millimeter-wave bands, and coordinate potential 6G spectrum to ensure long-term capacity. Consider fee reductions or exemptions to encourage ecosystem development. |
| **Strengthen Infrastructure Development:** | Build next-generation connectivity tailored to Mobile AI requirements - ultra-low deterministic latency, massive uplink capability, always-on connections, and ultra-high reliability. Accelerate the implementation of key 5G-Advanced and 6G technologies, such as AI-Native and Agent Communication. Lower regulatory barriers for continual upgrade and iteration. Increase edge node deployment and support distributed inference and federated learning. Build integrated orchestration capabilities for network, computing power, and data, supported by an open platform ecosystem. |
| **Activate Data as a Core Production Factor:** | Promote secure cross-domain and cross-border data flow. Open access to high-quality, scenario-specific datasets while establishing clear frameworks for data ownership and benefit sharing. Encourage privacy-preserving mechanisms such as data usable but not visible to build trust and increase willingness to share data. Develop industry-specific datasets and data-asset markets to drive new combined "network + computing + data" business models. |
| **Drive a Virtuous Cycle of Industry Innovation:** | Encourage applications to open APIs for cross-APP intent understanding and task orchestration, unleashing the full potential of agent technologies. Use targeted subsidies and funds to stimulate demand for Mobile AI terminals. Launch Mobile AI pilot programs in key verticals to catalyze ecosystem growth. Promote inclusive, cloud-assisted intelligent terminals to scale lightweight, low-cost AI experiences. Guide precise industry investment, build multi-stakeholder value-sharing mechanism, accelerate technology iteration, and strengthen talent development across "industry + communications + AI". |

Only through the combined power of policy, technology, and commerce, under the guiding principle of "Ethical AI with Reliable Connectivity", can the disruptive potential of Mobile AI be transformed into shared digital prosperity. Let us work together to build an open, innovative, and responsible Mobile AI ecosystem - and open a new chapter in the integrated evolution of mobile communications and artificial intelligence.



**This white paper was developed and published under the joint initiative and coordination of GSMA and GTI .**

### *Contributing Organizations*

We extend our sincere appreciation to the following organizations for their substantial contributions to the research, drafting, and review of this white paper:

**China Mobile | China Telecom | China Unicom | China Broadnet | Huawei | ZTE**
**Qualcomm | Nokia | Omdia | Honor | e& UAE | Droi | Showmac | Veezhen Consulting**

### *Supporting Organizations*

We also express our deep gratitude to the following organizations for their valuable support, including the provision of technical insights and industry expertise during the development of this document (listed alphabetically):

**AGrandTech | Airtel | HKT | Keysight | KuaiShangYun | NXCLOUD | OPPO | Rohde & Schwarz**
**Tongxin Micro | UNISOC | vivo | Xiaomi | YTL | Z.AI | Zain KSA | Zain Kuwait**

**1 Angel Lane**
**London**
**EC4R 3AB**
**United Kingdom**

Tel:+44 (0)20 7356 0600
Fax:+44 (0)20 7356 0601