GTI

Mobile AI

Application Experience Metrics

June 2026

Contents

Preface

The in-depth convergence of mobile communications and artificial intelligence (AI) is propelling the global digital economy into a new era of Mobile AI, serving as a core driving force for global digital and intelligent transformation. Previously, during MWC26 in Barcelona, GTI and GSMA, together with global industry partners, officially released the Mobile AI White Paper. This milestone publication systematically expounded on the core essence and systemic framework of Mobile AI for the first time, innovatively introducing a "Three-Layer, Four-Dimension" core architecture. It has outlined a development path for the deep convergence of mobile communications and AI, successfully fostering a consensus for global industrial development.

At present, AI technologies have made comprehensive breakthroughs. Emerging service forms such as multi-modal interaction, AI agents, and embodied AI are rapidly surfacing, profoundly reshaping the human-machine interaction mode and industry service paradigm. Concurrently, 5G-Advanced (5G-A) is accelerating commercial use, and 6G technical research and standardization are advancing rapidly. "Endogenous intelligence" has been established as the core technical direction for 6G. Major global economies are accelerating the allocation and implementation of 6G spectrum planning, solidifying the foundation for industry growth. The convergence of networks and intelligence exhibits a trend of "Network for AI + AI for Network". As a new digital resource, tokens are driving the industry to transform from traditional "traffic monetization" to "token monetization". Mobile AI has transitioned from the phase of technical exploration and proof-of-concept (PoC) into a critical period of large-scale commercialization across multiple scenarios. Against this backdrop, user experience has become the core indicator for measuring the quality of Mobile AI services, as well as a pivotal factor supporting sustainable and high-quality industry development. However, traditional application experience evaluation metrics, originally designed for voice, video, and other conventional services, fail to adapt to the unique characteristics of Mobile AI operations. The industry has yet to establish a unified, scientific, and quantifiable evaluation system for Mobile AI application experience. This gap leaves experience across different vendors and applications incomparable, makes it difficult to precisely map network resource optimization to service experience enhancements, and causes industry development to become increasingly fragmented.

As a significant continuation of the Mobile AI series, this white paper focuses on the Mobile AI application experience metrics, a core issue of global industry. It systematically constructs a hierarchical Mobile AI experience evaluation framework spanning the user perception layer, quality characteristic layer, and network performance layer. Furthermore, it defines key experience indicators and quantitative methods for typical service scenarios, clarifies the quantitative mapping mechanism between service experience and network and computing capabilities, and ultimately forms a complete set of methods for evaluating the quality of Mobile AI application experience. Grounded in a perspective of global industry collaboration, this white paper consolidates technical consensus among all parties in the industry chain. It aims to provide a unified experience metric benchmark and technical implementation guidelines for global Mobile AI application developers, network operators, equipment manufacturers, and standards organizations. By doing so, it drives a paradigm shift in network construction and optimization from "network indicator-driven" to "user experience-driven", accelerates the implementation of the Mobile AI application experience metrics, and help build an open and win-win global Mobile AI industry ecosystem.

GTI

Chapter 1:

Research Background and Industry Status

The deep convergence of mobile communications and AI is ushering in a new era of Mobile AI. As 5G-A undergoes deepened commercial adoption and 6G technical research advances globally, Mobile AI applications are leaping from single-function tools to all-scenario intelligent agents. Emerging business forms, such as multimodal interaction, AI agents, and embodied AI, are accelerating their market entry. This rapid evolution creates an urgent need for refined and standardized evaluations of service experience. This chapter systematically describes the background and necessity of the research on Mobile AI application experience metrics from three dimensions: industry development trends, industry status and core challenges, and research objectives and industry value, laying a foundation for the construction of subsequent evaluation systems.

1.1 Mobile AI Industry Development Trends

Network and intelligence convergence is now central to global digital economy. All-round breakthroughs in AI technologies, combined with accelerated 6G standardization, are driving network-intelligence convergence from technical exploration toward in-depth implementation. This convergence is systematically reshaping the technical frameworks, terminal form factors, and application ecosystems of the mobile communications industry. Consequently, it is propelling the Mobile AI industry to enter a phase of explosive growth marked by large-scale commercialization across multiple scenarios and driving the digital transformation of global industries.

Explosive AI Development

The AI industry is witnessing explosive development in terminals, technologies, applications, and business models.

- Terminals: Intelligent terminals are proliferating at an accelerated pace. According to Omdia, AI-capable smartphones will account for 54% of total smartphone shipments by 2028, reflecting a compound annual growth rate (CAGR) of 63% from 2023 to 2028. Beyond smartphones, shipments of smart wearables, such as AI glasses, are also increasing.

54%

Market share

(AI phones / total smartphones, 2028)

63%

Growth rate

CAGR (2023–2028)

- Technologies: Multimodal intelligent technologies are evolving from single-modal processing to full-modal convergence. Seamless interworking and joint modeling are achieved for texts, images, audios and videos, and multi-dimensional sensor data. A closed-loop capability covering the entire perception-understanding-reasoning-generation process has fundamentally taken shape. Furthermore, the device-edge-network-cloud collaborative architecture continues to optimize, while lightweight foundation models and on-device deployment technologies are becoming increasingly mature.
- Applications: AI agents are being widely adopted. Capable of environmental perception, autonomous decision-making, and closed-loop execution, these agents have enabled an autonomous service paradigm of "one instruction, full-process delivery." They have been commercially verified across multiple vertical fields, including automated data processing, intelligent e-commerce operations, enterprise office collaboration, and industrial production assistance.
- Business models: The token economy system is poised for takeoff. As foundation models and AI agents are rapidly gaining popularity, the core resources consumed by users have evolved from traditional network traffic to a comprehensive resource system that includes network connectivity, computing services, and AI capabilities. As a unified unit of measurement for foundation model inference and intelligent services, tokens are emerging as a new type of digital resource and core operational asset, following traffic, and are used to measure the value of computing power, models,

data, and services in a unified manner and facilitate efficient circulation of these resources. This will drive the industry to transition from traditional traffic monetization to token monetization, and upgrade communication services from basic connectivity to comprehensive intelligent services.

Accelerated Network Evolution

In 2026, the global mobile communications industry will enter a critical period for large-scale commercial use of 5G-A and accelerated implementation of 6G vision. Globally, over 120 operators across more than 45 countries and regions have initiated 5G-A technical verification or pre-commercial network deployments. AI applications are penetrating into the industry from the consumer side. Intelligent terminals, embodied AI robots, and Internet of Vehicles (IoV) are becoming increasingly popular, driving a shift in mobile network requirements from being predominantly downlink-centric to emphasizing both uplink and downlink capabilities. In addition, networks are evolving from the traditional "best-effort" mode toward a paradigm capable of delivering "deterministic" experience guarantees, providing predictable, measurable, and guaranteed differentiated services for diverse operations.

2025	2026	2028	2029	2030
3GPP launches 6G standardization	R19 frozen; 120+ operators verify 5G-A	AI phones reach 54% market share	R21 frozen; 6G pre-commercial begins	6G large-scale commercial deployment

In addition, 5G-A core standards have been commercially anchored. As a key milestone of 5G-A, 3GPP Release 19 (R19) specifications have been officially frozen. This not only deepens cutting-edge capabilities such as integrated sensing and communication (ISAC), non-terrestrial network (NTN), and green endogenous intelligence, but also takes the lead in completing the "AI-ready" architecture design at the standards level. On another front, global 6G technical research and industrial layout have entered a substantial stage. 3GPP officially launched 6G standardization in 2025. In Release 21 (R21), "endogenous intelligence" is about to be identified as a core technical direction, targeting a release freeze by March 2029.

Countries around the globe are accelerating the planning and allocation of experimental spectrums for 6G. Major economies such as China, the European Union, and the United States have released 6G spectrum planning schemes, specifying the coordinated deployment of mid- and low-band and mmWave bands. The Ministry of Industry and Information Technology (MIIT) of China has approved the use of the 6 GHz band for 6G experimental frequencies for the IMT-2030 (6G) Promotion Group, making China the first nation globally to approve 6G experimental frequencies.

Currently, the upstream and downstream of the industry chain are conducting joint research and technical verification focusing on core areas such as air interface technologies, native AI network architecture, and terminal prototypes. Operators, equipment vendors, and chip enterprises are collaborating to promote the implementation of key 6G technologies. A global industrial consensus has fundamentally taken shape. The first batch of 6G networks is expected to start pre-commercial deployment by 2029 and enter the large-scale commercial use around 2030. This will provide a network foundation with native experience and endogenous intelligence to support the next-generation evolution of Mobile AI.

Network-Intelligence Convergence and Network Form Reconstruction

From the perspective of the global telecom industry, Mobile AI is evolving from "single-model capabilities" to "AI-native networks + AI agent ecosystems." The bidirectional convergence of networks and AI is transforming from an "additional function" to "ubiquitous intelligence" that runs through all network domains and levels, forming a comprehensive, native closed loop of "networks supporting intelligence, and intelligence optimizing networks".

Network for AI: Evolving from "Centralized Models" to "Intelligent Digital Networking"

The deployment of AI applications for mobile networks is breaking the limitations of centralized foundation model inference in traditional data centers. The AI applications are rapidly advancing toward distributed autonomous entities across diverse scenarios, including terminals, vehicles, and cities.

- **Network element (NE) capability support:** Leveraging 5G SA and future AI-native 6G, networks can provide differentiated, low-latency, and more energy-efficient intelligent connections for massive devices and multi-modal scenarios through deterministic transmission assurance capabilities such as network slicing, edge computing, and communication-computing convergence.
- **Infrastructure reshaping:** 5G-A is accelerating its evolution to 6G, and is deeply intertwined with AI to build a new infrastructure of "intelligent digital networking." It is oriented to digital twins, real-time translation, fraud detection, and mission-critical communications. It breaks the bottleneck of traditional single-terminal computing power or applications, and promotes the transformation of Mobile AI service experience to the network side, where AI traffic and intelligent agent interaction modes are accurately perceived, and real-time response and endogenous scheduling are performed on the network.

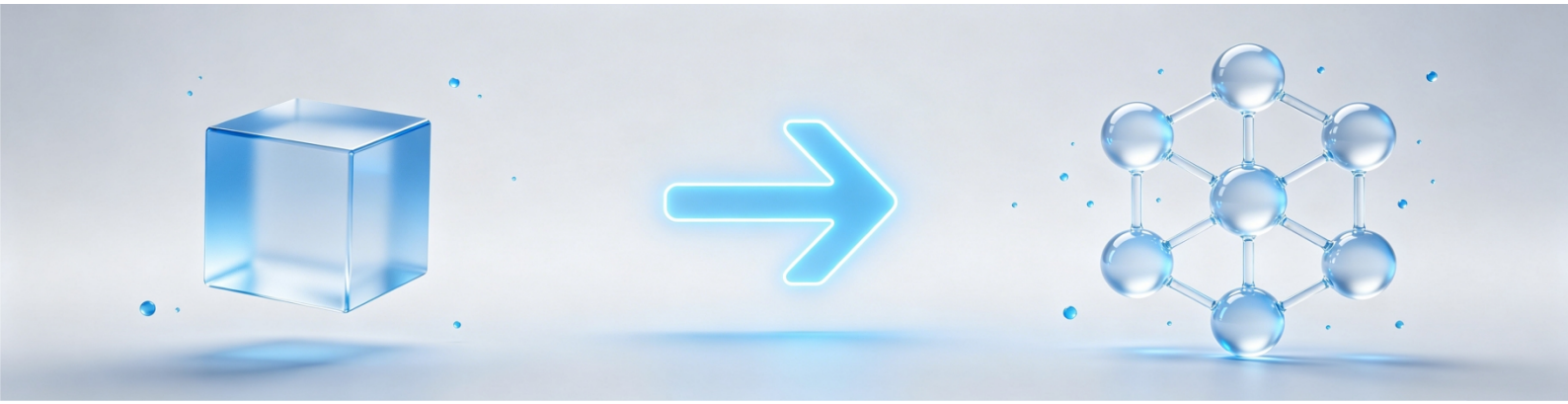
AI for Network:
Leaping from "Single-Point Model" to "Autonomous Networks"

The AI form inside the network is being upgraded from a single-point functional model to an autonomous network centered on the large language model (LLM), retrieval augmented generation (RAG), and agentic AI. By deeply applying AI across the entire lifecycle of communication networks, it achieves automation and intelligence in network planning, operation and maintenance (O&M), optimization, and operations, thereby elevating network efficiency and autonomy.

- **Zero-touch autonomous management:** In terms of service forms, generative AI and agentic AI will transform telecom networks from "passive infrastructure" to "intelligent partners" that can engage in conversations. Operations personnel and developers can express service objectives and O&M intents using natural language. Through conversational AI, intent-driven management, and multi-agent collaboration, the network can achieve a closed-loop automation that is "conversational, perceptible, inferable, and executable" across the planning, O&M, optimization, and operations of the entire service lifecycle. This supports the continuous evolution of networks towards truly zero-touch autonomous networks.
- **Telecom domain adaptation:** To ensure the reliability and service value of AI in complex telecom operations environments, the industry is accelerating the adaptation of LLMs to the telecom domain. This includes constructing specialized corpora and evaluation datasets, alongside introducing multi-dimensional metrics to systematically assess the output quality of generative AI.

Hierarchical function architecture:
AI agents broadly embedded as a universal technology

In terms of the network architecture, AI agents are regarded as a universal technology and seamlessly embedded into various functional layers across the core network, intent management, and business support system (BSS). Enabled by mechanisms such as dynamic network slicing, predictive O&M, and multi-core network connection/agent-to-agent (MCP/A2A), AI is gradually evolving mobile networks into a truly autonomous, explainable, and ecosystem-oriented intelligent foundation, with rigorous security, robustness, and trustworthiness ensured.



Bidirectional Transformation of Terminal Form Factors and Traffic Paradigms

- **Terminal-workflow synergy:** As diverse terminals such as embodied AI robots, AI-powered smartphones, and wearable AI agents are emerging, a hierarchical intelligent terminal system is gradually established. (AI agents capable of performing closed-loop tasks autonomously at level 3 are becoming mainstream.)
- **Redefinition of traffic and experience metrics:** Token streams and intelligent agent interaction processes are transforming into a new type of network traffic. This poses new challenges to traditional network experience metrics, urging the industry to accelerate the introduction of an indicator evaluation system oriented to agent workflows in future network standards and experience evaluation.
- **Multi-scenario commercial implementation:** Mobile AI has been fully integrated into three core fields, fostering a booming momentum of multi-scenario commercial deployment. These core fields include life, production, and governance: In the field of life, it has given rise to new consumption experiences such as immersive interaction and personalized services. In the field of production, it has promoted industry upgrade models such as flexible production and intelligent inspection. In the field of governance, it has built a public service system featuring refined management and intelligent response.

1.2 Industry Status and Core Challenges

As Mobile AI applications become increasingly popular and service forms continue to diversify, global users are placing higher demands on the experience of AI services. However, due to physical limitations such as terminal computing power, power consumption, implementation complexity, and form-factor size, future AI application scenarios will rely heavily on the network or cloud to provide LLM inference services to the device side. In this process, network conditions such as radio channel quality and IP-layer latency and jitter, as well as the deployment mode of LLMs, will directly determine the AI service experience on the device side.

What's more, the deployment location of AI inference services on smart terminals will dynamically change in the future. The system needs to dynamically optimize and schedule models based on real-time conditions: for instance, offloading inference tasks to the network side when terminal battery levels are low to reduce terminal power consumption, while automatically switching to light-weight, on-device models in poor signal or off-grid environments to guarantee basic experience. This dynamic mechanism of cloud-edge-device synergy makes it essential to scientifically evaluate the impact of user experience and network conditions.

Currently, the industry has yet to establish a unified experience evaluation system that adapts to the characteristics of Mobile AI services. Traditional network evaluation methods are significantly disconnected from service requirements. For a long time, the network evaluation system in the communications industry has been built around pipe key performance indicators (KPIs) such as throughput, latency, and packet loss rate, aiming to meet the requirements of traditional services such as voice and video. However, Mobile AI services exhibit unique features, such as multimodal interaction, token streaming, and task-oriented execution. The overall user experience depends not only on the underlying transmission performance, but also on service-layer indicators such as the first token latency, output smoothness, multimodal content synchronization, and task completion accuracy. Under the same E2E network latency, different first token response speeds and generation rates will lead to entirely distinct user perceptions. Traditional KPIs cannot effectively quantify these experience differences.

Furthermore, there is no unified experience measurement benchmark globally. Different countries and vendors use their own evaluation dimensions, weights, and evaluation criteria, resulting in a lack of comparability and mutual recognition of evaluation results. This leads to fragmented industry development. As a result, it is difficult for operators to accurately identify experience bottlenecks of AI services and perform targeted resource scheduling. In addition, because the quantitative mapping mechanism among Mobile AI experience and network, computing power, and terminal capabilities has not been established, the network side cannot provide differentiated quality of service (QoS) guarantees based on diverse AI service requirements. This not only makes it difficult to achieve an optimal balance between network resource utilization and service experience, but also fails to provide a clear experience orientation for the design and standardization of the global 6G network architecture. Therefore,

reconstructing an experience evaluation system tailored for Mobile AI is of great importance for future network design and optimization.

1.3 Research Objectives and Industry Value

This white paper, as an important continuation of the Mobile AI series, aims to address the core issues in the global Mobile AI industry, such as the lack of experience evaluation metrics and the disconnect between experience and network capabilities. It aims to formulate a scientific, unified, and actionable global quality of experience (QoE) evaluation system for Mobile AI applications, provide common technical support for worldwide industrial development, and foster collaborative, healthy growth across the global Mobile AI industry.

The core research objective of this white paper is to (1) establish a hierarchical Mobile AI experience evaluation system that covers the user perception layer, quality characteristic layer, and network system layer, (2) define key experience indicators and quantitative methods for typical service scenarios such as multimodal interaction, AI agents, and embodied AI, (3) clarify the mapping mechanisms between service experience and network and computing capabilities, and (4) form a complete set of methods and grading metrics for evaluating the quality of Mobile AI application experience. Through this system, the gap between user experience and underlying network performance will be bridged, and the network construction and optimization model will shift from "network indicator-driven" to "user experience-driven". This will spur the large-scale development of AI applications on 5G-A networks, enhance users' perception of intelligence, and provide guidance for the in-depth convergence of global networks and intelligence in the 6G era. This system can offer a universal experience metric reference for Mobile AI application developers and provide experience evaluation tools and optimization approaches for network operators, thereby facilitating the collaborative progress and prosperity of global network-intelligence convergence.

GTI

Chapter 2:

Framework for Evaluating Mobile AI Application Experience

2.1 Research Scope

Foundation models are moving on the road to a new phase of AI: artificial general intelligence, where agents are pivotal to convert model capabilities into real value. Such AI applications would not be possible if the intelligent closed-loop is not formed among sensing, thinking, and acting, which hinge on three core systems: a sensing system that captures multidimensional environmental information, a thinking system that converts the fragmented data into structured knowledge and performs inference and planning, and an acting system that interacts with the environment to exert control over the physical or virtual world.

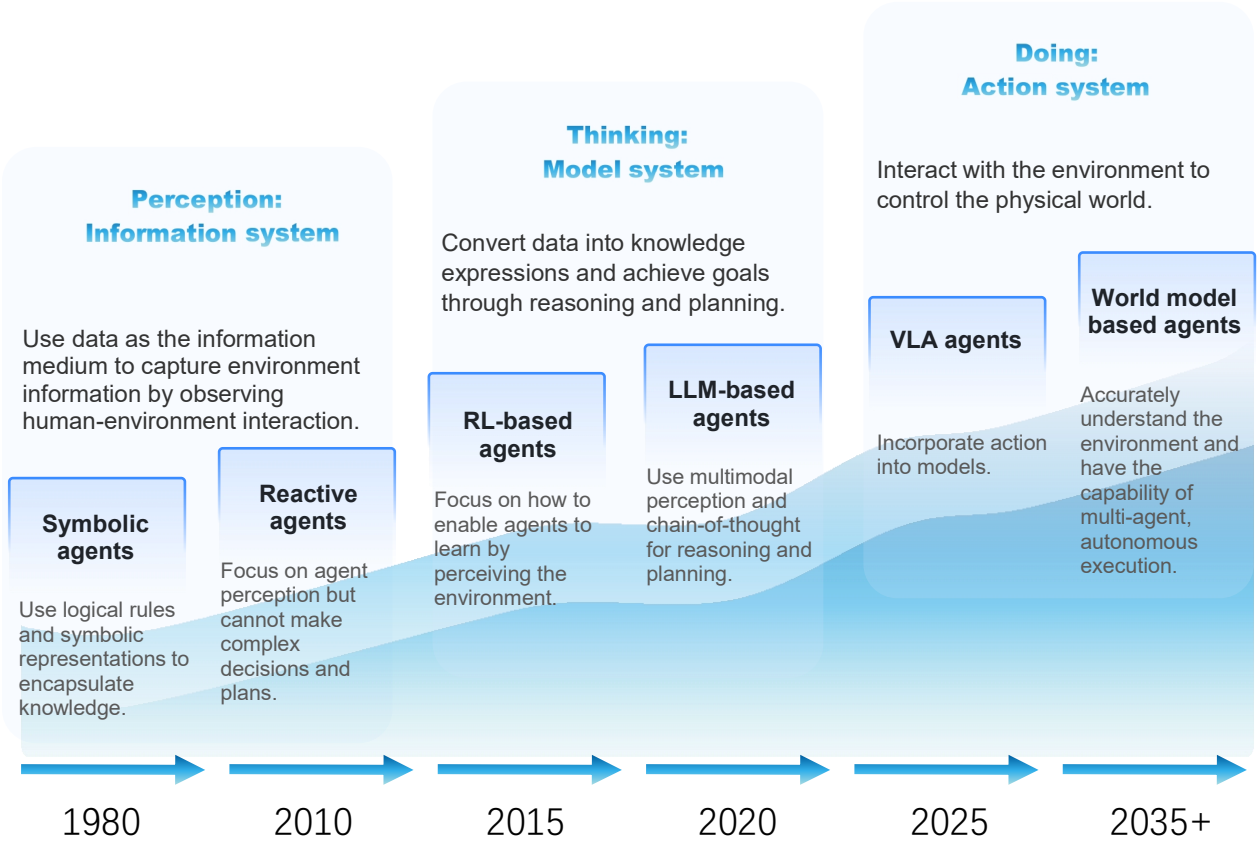


Figure 2-1 Agent Evolution Path

This capability mix maps to the three core use cases of Mobile AI services and places new requirements on networks and network evaluation.

First: multimodal interaction. Users exchange text, audio, and visual information with AI simultaneously or consecutively. This is a basic Mobile AI service that requires both single-modal data transmission and high-precision chronological alignment and semantic consistency between modalities to ensure user experience. This underscores the necessity to shift the evaluation focus towards cross-modal recognition accuracy, response synergy, and E2E coherent interaction.

- **Second: AI agents.** Foundation models are augmented with new sensing, inference, and acting capabilities, moving from reactive support to proactive solutions. The interaction intensity can be categorized as strong, moderate, or weak, based on latency tolerance. Strong interaction requires extremely low latency, while moderate interaction requires human-like response for services. Weak interaction is not time-sensitive to support such features as content generation. Accurate intent understanding, rational task breakdown, and cross-application smoothness are major factors in evaluating user experience.
- **Third:** embodied AI. Physical entities like robots and autonomous vehicles interact with environments. This is the ultimate form of AI, fully capable of sensing, thinking, and acting, which

are mapped to vastly different network capabilities. For example, robots prioritize ultra-low latency and ultra-high reliability for remote control, while also needing high-precision timing synchronization for robot collaboration, as this will enable them to leverage model size and transmission frequency to execute tasks. The success rate of sensing, decision, and action and accurate physical interaction for specific tasks are vital to evaluating user experience.

These dynamics call for a break from the evaluation system centered on KPIs such as throughput and basic latency, as networks are evolving from passive pipes to intelligent hubs. It is imperative to adapt the evaluation framework to Mobile AI services by defining clear boundaries and indicators, so as to set up the groundwork for 6G network design and experience optimization.

Considering the development of AI foundation models and their adoption in the industry, this paper studies the quality evaluation of user experience for real-time interactive and task-oriented applications. Research on embodied AI is still underway and will be disclosed as the results are mature.

2.2 Hierarchical Architecture for Experience Evaluation

This document proposes a progressive three-layer QKK architecture for experience evaluation: QoE layer, key quality indicator (KQI) layer, and network key performance indicator (KPI) layer. This architecture fully associates user experience with network parameters, offering an effective model to standardize and accurately quantify user experience and characterize underlying indicators for Mobile AI services.

QoE layer: multidimensional comprehensive satisfaction

The architecture places QoE at the top layer to directly reflect how satisfied users are with multimodal AI systems. This layer uses structured scoring and objective regression model calibration to reflect the integrity and diversity of experience dimensions, which focus on three core aspects:

- **Availability:** It is the cornerstone to user experience, with a focus on function integrity, operation stability, and content presentation reliability. It indicates the accessibility and frictionless experience of services.
- **Interactivity:** It reflects a core advantage of multimodal AI systems, focusing on the naturalness, coherence, and efficiency of multimodal interactions like as audio, vision, and text.
- **Intelligence:** It reflects high-level capabilities for AI systems to understand intents and provide accurate services autonomously to meet personalized user needs.

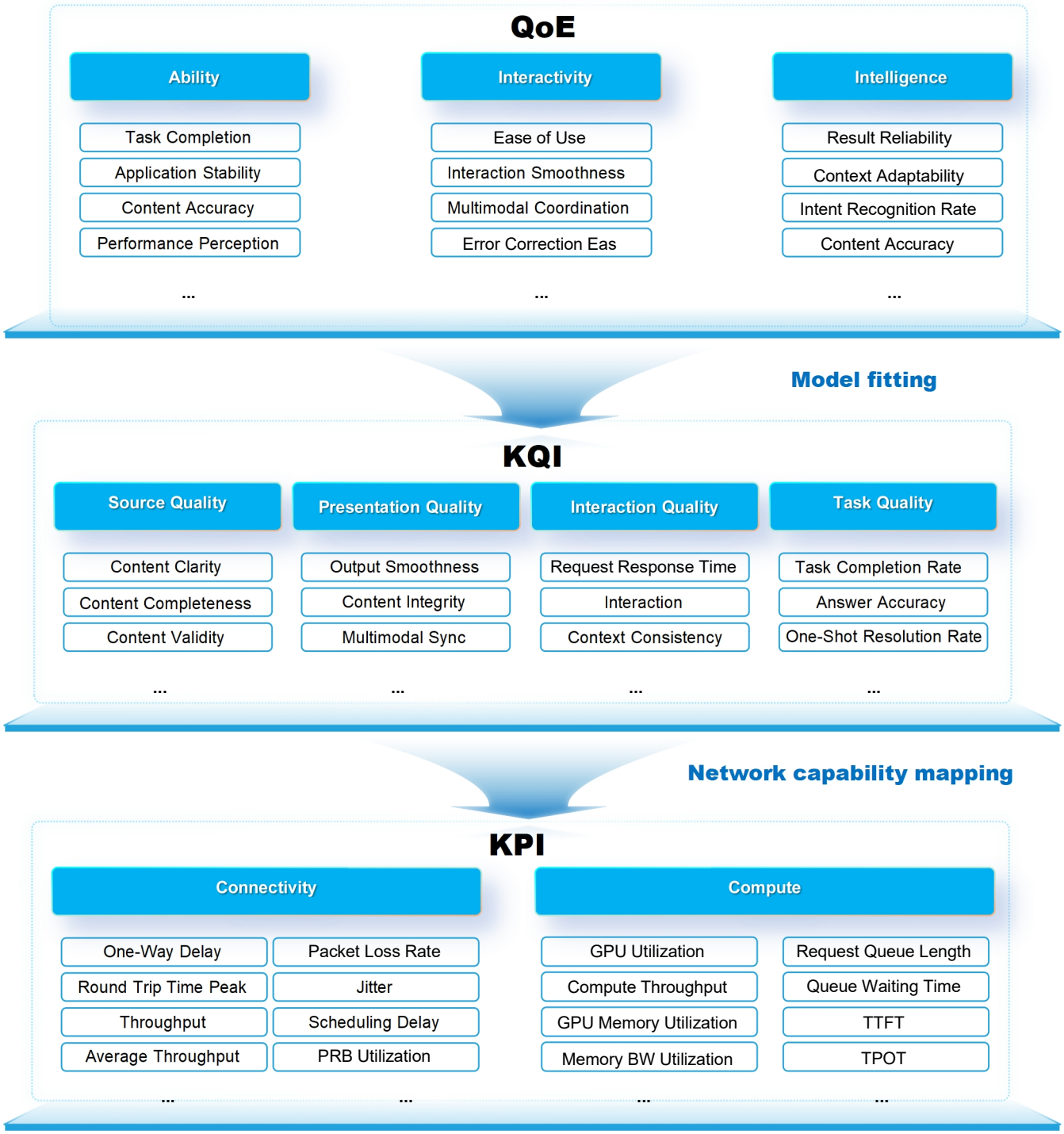


Figure 2-2 User AI Experience Dimensions

Scoring weights w1, w2, and w3 are applied to experience evaluation in line with both system factors and scenario-specific factors.

System factors are the attributes and characteristics that are decisive to the quality of application or services. They include the AI systems that interact with users, entire hardware and software settings, and user-perceived system design. The factors may vary with multimodal AI systems.

Scenario factors are related to environment variables or context parameters. They factor into the signals exposed to users, goals of using specific systems, and user expectations on applications, all of which impact the quality of user experience. The user-perceived quality measures how closely a service meets

user expectations, which largely depends on the situation in which the service is used, and therefore cannot be ignored when assessing subjective experiences.

KQI layer: quantitative mapping of service characteristics

KQI links the QoE layer to the network/system layer, abstracting user experience into service indicators that can be quantified and measured to accurately reflect AI token transmission and autonomous task execution.

- **Source media quality:** It measures the media quality input to multimodal AI systems without being subject to network damage, including both the clarity and effectiveness of images and audios. The image quality is determined by a mix of resolutions, frame rates, bit rates, and pixel depths, while the audio quality factors include bit rates.
- **Presentation quality:** It measures how multimodal presentation on devices affects user experience during AI interaction, focusing on output smoothness and coordination. It can represent experience impairment caused by network transmission and is closely related to response integrity and smoothness. The indicators include image and audio freezing rates, modal asynchronization rates, and asynchronization duration. For a multimodal system, smooth and coordinative presentation to users is a key factor in user experience.

For models with streaming generation, the token generation rates (tokens/s), output smoothness, and multimodal alignment are the key evaluation targets to avert impaired immersion due to freezing frames.

- **Interaction quality:** It represents the real-timeliness and engagement of AI multimodal interaction, indicating the agility of response links in terms of responsiveness, communication efficiency, and interaction effectiveness. It measures how the responsiveness and interaction of AI multimodal services affects user experience. The interaction quality affects how immediately the expectations of users on services are fulfilled.

The core indicators include time to first token (TTFT), number of interaction turns, and E2E response latency. The TTFT measures the time between a user sending a request and the model returning the first token. An interaction turn is a single instance consisting of one user input and one application output. A complete turn starts when a user sends an input and stops when the multi-modal AI application displays all output or when the display is aborted. A turn is considered incomplete when the output is not fully received due to timeout or other causes. The E2E response latency is the time elapsed between the end of user input and the commencement of multimodal AI output.

- **Task quality:** It measures the effectiveness and efficiency of task execution for users during AI multimodal interaction. The task success rate and efficiency are designed to evaluate the probability of completion success and the time required for scheduled tasks, which are core to system practicality. The task quality also covers experience discontinuity due to interruptions between the user and AI application when they are trying to understand each other. For AI agents and embodied AI, the task success rate, result accuracy, and automation rate are prioritized to better measure the value of transition from interaction to execution.

Network KPI layer: underlying support for full-stack capabilities

This layer covers the physical foundation for experience delivery. It includes common network indicators and also integrates computing and AI-native network characteristics to provide all-round experience support.

- **Radio access and transmission indicators: These touch upon channel quality (via RSRP, SINR, etc.) and E2E latency (transmission, queuing, and processing), jitter, and reliability, which directly support for interaction quality at the KQI layer.**
- **Computing and resource indicators:** These include system-level KPIs like GPU usage, memory usage, and inference queue lengths, reflecting the hardware responsiveness of model inference.
- **System O&M indicators:** For service scheduling, they include the QoS policy accuracy for burst AI token flows, and the SLA fulfillment rate of network slices. For autonomous tasks, fault locating duration, automatic handling rate, and intent-driven service stability are designed to adapt to the closed-loop of sensing, searching, thinking, and acting.

This model measures how network quality parameters affect user experience through the KQI layer, offering a practical basis for optimizing multi-modal AI systems.

2.3 Indicator Definition and Mapping

Holistic systematic verification is required to reliably adapt the hierarchical mapping model to the full category of multimodal AI applications, including chatbot, visual processing, agent tasks, and robot control. Cross-validation in different scenarios help eliminate accidental impact of parameter changes at each network layer in a certain scenario on user experience, which is crucial to build statistically significant mapping relationships.

This mapping model provides accurate guidance for optimizing multimodal AI systems.

- **Fault demarcation and quick locating:** When a QoE indicator worsens, the mapping relationship allows the fault to be quickly mapped to the KPIs at network/system layer, setting up for accurate locating.
- **Optimization prediction and value alignment:** The benefits of an underlying improvement to user experience are evaluated in advance to ensure that the optimization always matches user needs. This offers an effective systematic methodology to convert technical indicators into user value in the course of system iteration.

Impact of Key Parameters on Experience Quality

Based on experiment data, indicative underlying parameters are selected to explain their impact mechanism on user experience (availability, interactivity, and intelligence).

Audio source quality parameters determine the immersive-ness of multimodal interaction, and are essential to building in-depth interaction experience in all scenarios.

- **Audio bit rate:** Indicates the amount of audio data collected, coded, and uploaded by an AI multimodal client per unit time, in the unit of kbps.

Visual source quality parameters determine the definition and stability of multimodal interaction, and are the core to ensuring basic user experience.

- **Frame rate (FPS):** As a core indicator of video streaming media, it directly reflects image coherence. Fluctuating or low frame rates are the common sources of unstable or freezing video call images, preventing users from experiencing stable services.
- **Bits per pixel (BPP), or pixel depth:** It reflects the trade-off between image compression quality and transmission efficiency. An excessively small BPP means oversized image compression and results in image distortion, color discontinuity, and detail loss. This clearly drags down content presentation quality and subjects users to downgraded experience.
- **Resolution:** It is a basic metric of visual quality, directly indicating content fineness. Low resolution is a main cause of image blurring, preventing users from obtaining clear visual information. This leads to poor satisfaction with content quality and system stability.

Interaction timeliness parameters are directly related to users' subjective judgment on system intelligence and agility.

- **E2E interaction response latency:** It is the time taken from the end of user input to the start of receiving output from the system. An increased latency delays a robot's execution of physical actions and path planning, leading to prolonged task completion. For chatbot services, this results in discontinuous interaction, significantly reducing user engagement and satisfaction.
- **TTFT:** It is the time taken from the time when a user starts a request to the time when the model returns the first token. A long TTFT delays the fetch of key feedback and interrupt interaction, creating a sense of lag or stutter for users. This would mean a severe damage to natural and smooth interaction.

Multimodal collaboration and continuity metrics mainly affect the consistency and reliability of multimodal systems.

- **Modal asynchronization rate:** It is unique to multimodal systems. A higher asynchronization rate (for example, more audio and video streams out of step, or unsynchronized voice commands and robot actions) results in irrelevant output or inaccurate execution. This interrupts the natural coherence of user interaction, dealing a direct blow to result reliability and physical behavior execution accuracy during AI interaction.
- **Video/Audio stalling rate:** It affects the continuity of visual and auditory modalities. Video stalling impairs an AI system's environmental sensing, including the ability to follow objects' movement trajectories, while audio freezing interrupts voice commands, reducing multimodal collaboration integrity and multi-sensor response consistency.

Task efficiency and intelligence parameters measure the holistic capabilities of an AI system's understanding and execution.

- **Success rate (SR):** This is a core metric for a system's function completion. A decrease in success rate directly points to the system's inability to fulfill user needs (navigation failure, incorrect command execution, etc.), reflecting the system's insufficient output accuracy and adaptation to the environment. This severely undermines user trust in the system.
- **Turns/cognitive efficiency:** It reflects a system's interaction efficiency and understanding capability. An increase in the number of turns required to complete a task means repeated clarification or correction by users. This is a direct indicative of weak understanding and complex interaction in the system, leading to a lower task completion efficiency.

The table below summarizes the mapping mechanism of key indicators.

Class	KQI	QoE Dimension	QoE Sub-Dimension	Mapping Basis/Impact Mechanism
Source media quality	Resolution	Availability	Application stability	A lower resolution decreases definition and inconsistent AI sensing performance, negatively affecting user experience stability.
	Frame rate	Availability	Application stability	Fluctuating frame rates cause unstable video call images, dealing a blow to system stability towards both users and AI systems.
	Frame rate	Interactivity	Interaction smoothness	A drop in significant frame rates will cause video freezing, directly reducing the smoothness of interaction.
	BPP	Availability	Application stability	A drop in BPP causes excessive image compression and poor visual sensing quality for AI systems, damaging application stability.
Presentation quality	Video stalling rate	Interactivity	Multimodal interactivity	Video stalling directly affects the integrity of multimodal collaboration (e.g., between visual and audio).
	Modal asynchronization rate	Interactivity	Multimodal interactivity	An increase in the modal asynchronization rate causes text, audio, and visual data to be out of pace, degrading multimodal experience.
	Modal asynchronization rate	Interactivity	Sensor response consistency	An increase in modal asynchronization rate causes multi-sensor data (like cameras and microphones) to be out of step, reducing response consistency.
	Modal asynchronization rate	Intelligence	Spatial understanding	An increase in modal asynchronization rate causes visual-audio inconsistency, directly affecting AI understanding of spatial structure.
Interaction quality	TTFT	Interactivity	Interaction feedback speed	A long TTFT delays a user's access to key outputs, leading to slow interaction feedback.
	E2E interaction response latency	Interactivity	Interaction smoothness	An increase in the number of turns interrupts interaction paces, affecting user experience.
	E2E interaction response latency	Intelligence	User engagement	A longer interaction latency increases the waiting time for users, reducing user engagement and satisfaction.
Task quality	Task success rate	Availability	Task completion rate	A decrease in the task success rate directly leads to users' failure to achieve expected goals, showing that the system cannot meet task requirements.
	Cognitive efficiency/turns	Intelligence	Result reliability	Repeated clarification during content generation directly reflects weak AI result reliability, increasing the interactions required to complete tasks.
	Cognitive efficiency/turns	Intelligence	Context memory	The need for users to repeat information from earlier conversations points to poor contextual memory, affecting interaction coherence.

Chapter 3:

Mobile AI Experience Quality Assessment Methods

3.1 Overall Methodology and Model Application Scope

The AI experience quality modeling process comprises four distinct stages: scenario construction, indicator extraction, user experiment, and quantitative modeling.

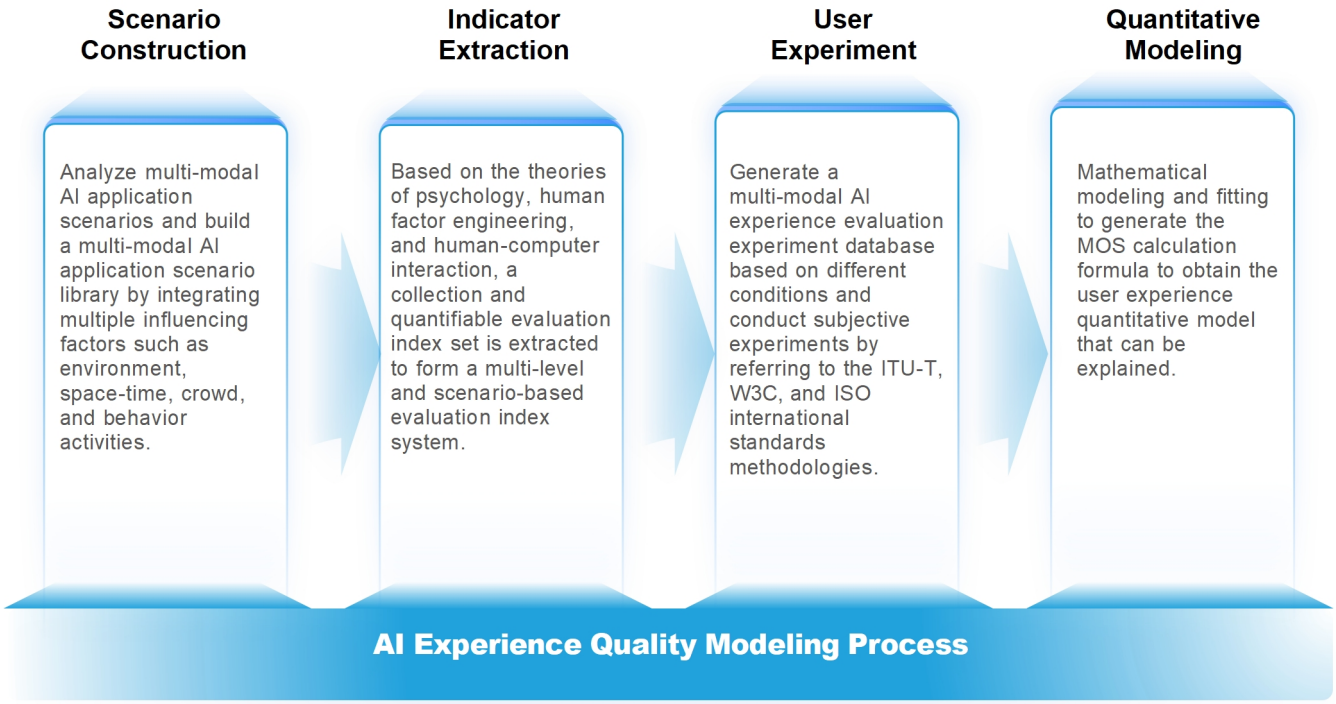


Figure 3-1 AI Experience Quality Modeling Process

AI multi-modal interactive applications span a wide range of usage scenarios such as online shopping, social networking, and travel navigation, enabling tasks such as retrieval, Q&A, and content generation. This study organizes user experiment test cases by modality combinations. These combinations are categorized into four types: image-text, text-audio, image-audio, and image-text-audio, tailored to the varying functions and tasks of AI multi-modal interactive applications across different scenarios. The applicable modality combinations, input/output modalities, application tasks, and use cases are detailed in the table below.

Modality Combination	Input/Output Modality	Application Task	Use Case Scope
Image-Text	Input: image, text Output: text, image	Image-text retrieval	(1) Text-to-image search (2) Image-to-text search
	Input: video, text Output: text, video	Video retrieval	(1) Text-to-video search (2) Video-to-text search
	Input: image, text Output: image, text	Static image Q&A	(1) Object recognition & explanation
			(2) Scene understanding
			(3) Sentiment analysis
			(4) Action explanation
			(5) Spatial relationship judgment
			(6) Common sense reasoning
			(7) Mathematical reasoning
			(8) Image-text consistency detection
	Input: video, text Output: video, text	Video Q&A	(1) Plot understanding
			(2) Character analysis
			(3) Sentiment analysis
			(4) Fact retrieval
			(5) Abstract reasoning
Text-Audio	Input: audio, text Output: text, audio	Text-audio retrieval	(1) Text-to-audio search
			(2) Audio-to-text search
Image-Audio	Input: audio-video Output: audio-video	Video anomaly detection	(1) Abnormal human behavior
			(2) Traffic anomaly event
			(3) Environmental anomaly
Image-Text-Audio	Input: audio-video, text Output: audio-video	Audio-video retrieval	(1) Video content understanding (2) Audio content understanding (3) Cross-media retrieval
	Input: audio-video Output: text	Audio-video text description generation	(1) Accurate description
			(2) Abstract description
			(3) Temporal description
	Input: audio-video, text Output: audio-video, text	Audio-video Q&A	(1) Audio-visual content understanding
			(2) Multi-modal interaction
			(3) Real-time information processing
			(4) Domain-specific Q&A

In the user experiment phase, either interactive or non-interactive experiments are selected based on the use cases outlined above. Participants provide subjective ratings reflecting their experience after exposure to specific stimulus combinations. To ensure experimental validity, stimulus presentation is controlled using appropriate sequencing methods, and validated scales are employed to measure participants' subjective experiences across multiple dimensions.

This study adopts the absolute category rating (ACR) method, as recommended in ITU-T P.910 Subjective video quality assessment methods for multimedia applications. In this category judgment method, only one stimulus combination is presented in each experiment. Participants provide independent ratings on an absolute scale, independent of any comparative context with other stimuli.

An experiment follows the single stimulus procedure when ACR is used for assessment. Participants provide immediate ratings after exposure to each stimulus, repeating this cycle until all stimuli have been evaluated. The inter-stimulus interval serves as the rating phase, where distractions are minimized to ensure focused assessment. The experimental stimulus sequence is illustrated in the following figure.

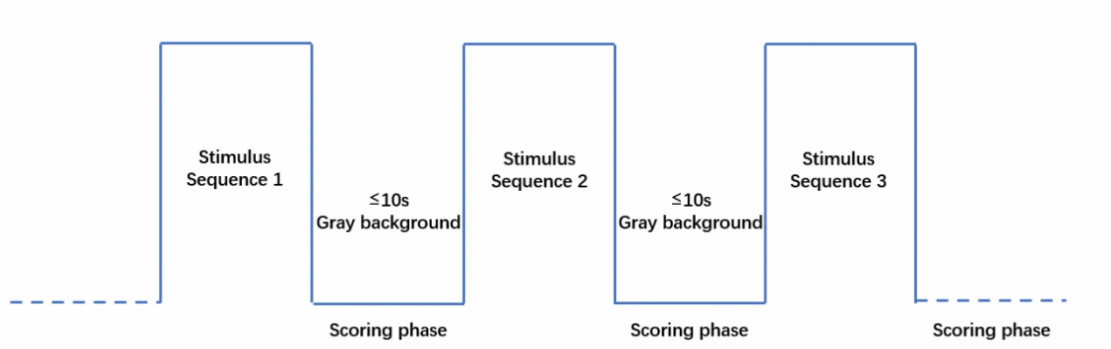


Figure 3-2 Experimental Stimulus Sequence and Rating Phase Diagram

Participants are the cornerstone of any assessment experiment. Experiment design must carefully define the participants' characteristics and their sample size. According to ITU-T P.910, the sample size refers to the number of independent ratings per stimulus. The selected sample size critically impacts the validity of the experimental outcomes. The standard recommends the following minimums: at least 4 participants for pre-experiments, 24 participants for controlled environment experiments, and 35 participants for non-controlled environments.

The experimental scale is determined by a trade-off between the experimental load and the sustained attention span of participants. ITU-T P.910 recommends limiting each participant's session to 1.5 hours to mitigate fatigue and boredom. If a longer duration is unavoidable, the design must incorporate frequent breaks and compensatory measures to offset these negative factors.

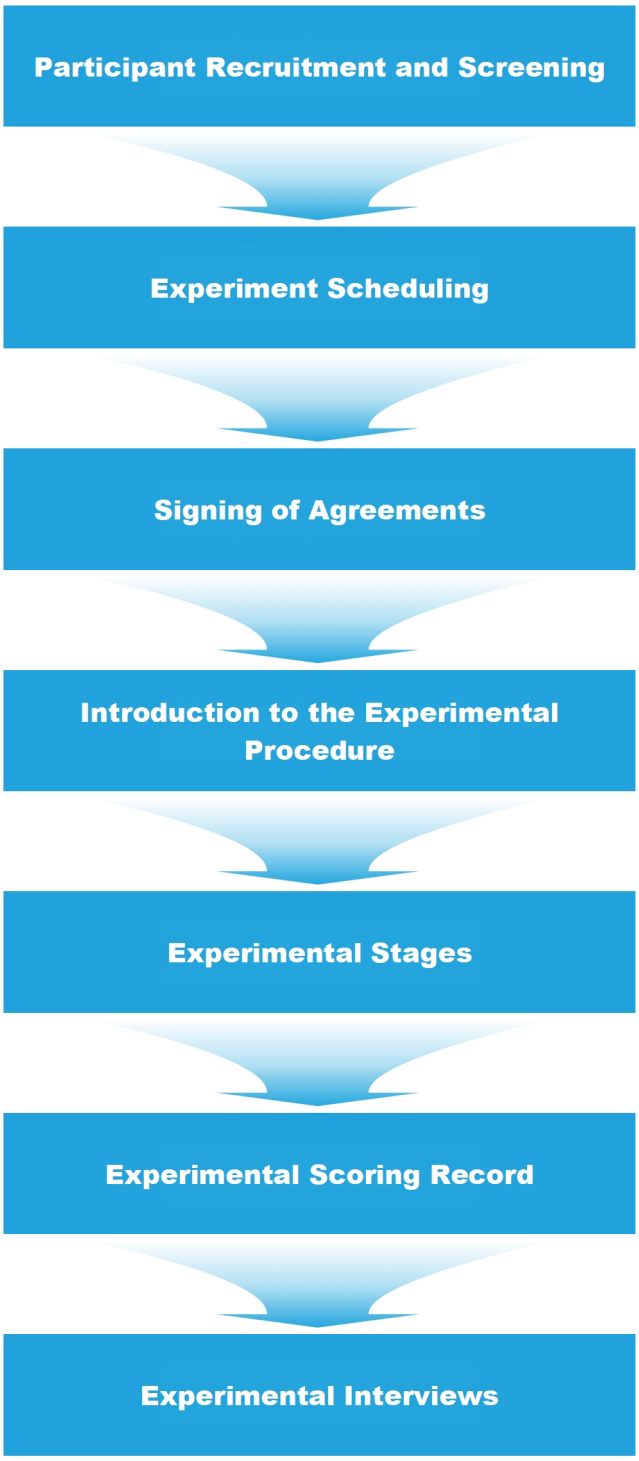


Figure 3-3 Assessment Experiment Procedure

Adhering to standardized scientific logic, an assessment experiment follows a systematic workflow comprising multiple stages, such as participant recruitment, agreement signing, instruction and training, experiment execution, and subsequent rating and recording. These stages are executed sequentially, as shown in the preceding figure.

3.2 Experiment Scheme and Scenario Library

As multi-modal AI becomes a pivotal partner in complex tasks and human-machine collaboration, there is an urgent need for a systematic and structured experience assessment framework. As multi-modal AI applications expand into sectors like lifestyle, education, and entertainment, the divergent nature of task goals and usage contexts creates highly disparate user expectations concerning AI capability boundaries, interaction efficiency, and error tolerance. Moreover, given the heavy reliance of mobile multi-modal AI on cloud computing power, user experience is strongly coupled with mobile network conditions. Consequently, constructing a multi-dimensional multi-modal AI scenario library is fundamental to advancing research in human-machine collaboration experience.

To address this, this study proposes an extensible scenario metadata structure that abstracts multi-modal AI scenarios into four core dimensions: physical space, user activities, AI functions, and target population. These are augmented by descriptive attributes such as modality combinations and device forms. This four-dimensional architecture provides a structured framework for designing typical scenario sets:

First, it generates a comprehensive set of representative scenarios by systematically cross-referencing the four core dimensions, and filtering out unreasonable or low-value combinations based on modality combinations and task risk levels.

Second, it selects typical scenario sets based on assessment requirements, covering representative application types and key tasks while balancing operability and extensibility. This creates a standardized benchmarking context for multi-modal AI user experience and quality assessment.

Finally, this scientifically classified scenario library acts as a standardized input for subsequent user subjective experiments, guaranteeing the integrity and validity of the experiment design.

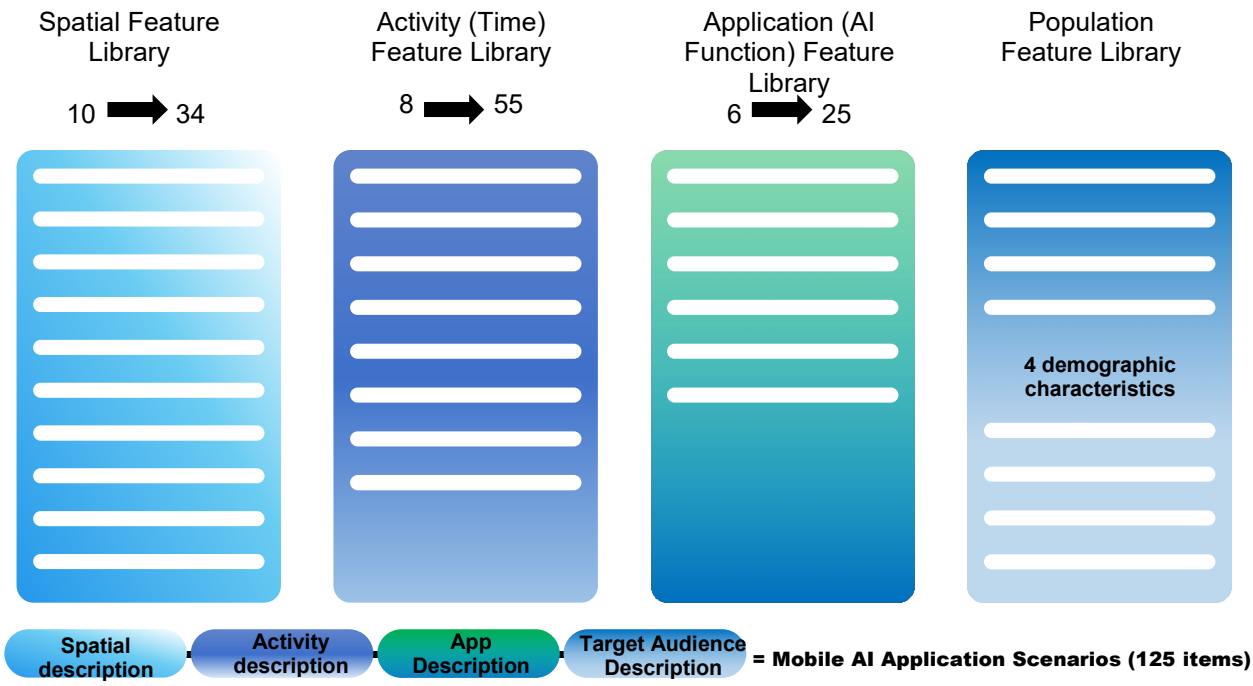


Figure 3-4 Figure 3-4 Experiment Scheme and Scenario Library

In constructing the scenario library, this study draws extensively upon authoritative domestic and international standards and statistical frameworks. These include the national standard Classification and coding of geographic information points of interest (GB/T 35648-2017), United Nations Statistics Division's International Classification of Activities for Time-Use Statistics 2016 (ICATUS 2016), and National Institute of Standards and Technology's AI Use Case Classification Framework. Through deep classification and aggregation of spatial activity characteristics and user behavior patterns, the following AI multi-modal usage scenario library is ultimately derived.

Primary Classification	Secondary Classification	Function/Task
(1) Productivity and Tools	(1) Office Assistant	Digital assistants (conversational agents), content synthesis (summarization/transcription), information retrieval, task suggestions
	(2) Knowledge and Information Management	Information retrieval, content integration (knowledge integration)
	(3) Writing and Content Generation	Content generation (text), content synthesis (translation/rewriting)
	(4) Enterprise Service Support	Digital assistant (partnering), recommendation (candidate answer, personalization, customer presentation)
	(5) Visual Creation	Content generation (painting), image analysis (style transfer)
	(6) Audio Creation	Content Generation (Audio)
(2) Healthcare and Health	(1) Diagnosis and Imaging Analysis	Image analysis, detection (lesion identification)
	(2) Clinical Decision Support	Prediction (disease risk/pre), decision-making (treatment plan)
	(3) Medical Consultation and Health Advice	Digital assistants (Q&A consultation), content integration (medical record organization)
	(4) Health Monitoring and Management	Monitoring (vital signs), personalization (exercise/health recommendations)
(3) Education and Knowledge	(1) Personalized Teaching	Personalization (learning content/learning progress), recommendation (learning path), decision-making (teaching strategy)
	(2) Evaluation and Testing	Decision (scoring), content synthesis (answer analysis)
	(3) Knowledge Retrieval and Question Answering	Information retrieval, content synthesis (generating summaries)
(4) Entertainment and Games	(1) Interactive Entertainment	Decision-making (NPC behavior), prediction (player behavior)
	(2) Multimedia Editing	Content integration (video editing), content generation (special effects)
	(3) Immersive Experience	Personalization (emotion and style), content generation (dialogue and virtual experience), content integration (multimodal immersive experience)
(5) Dialogue and Information Services	(1) Asynchronous session	Digital Assistant (Asynchronous Scenarios)
	(2) Real-time conversation	Content integration (translation/subtitles)
	(3) Intelligent Agents and Execution	Digital assistant (dialogue control), process automation (order placement/payment), device control (home appliances), personalization (user habits/home scenarios)
(6) Shopping and Lifestyle Facilities	(1) Information Aggregation and Recommendation	Information retrieval, content integration, recommendation
	(1) E-commerce shopping guide	Recommendation, Personalization
	(2) Virtual Experience	Image analysis (recognition), content generation (virtual effects)
	(3) Local life	Recommendation, Prediction (travel/dining trends)
(7) Travel and Transportation	(1) In-vehicle interaction	Digital Assistant (In-Car)
	(2) Public Transportation Travel Assistant	Decision-making, monitoring, forecasting, recommendation

3.3 User Pre-experiment Case

Based on the methodology described above, multiple representative scenarios are selected from the scenario library to conduct user subjective scoring experiments. The experiment samples are organized and analyzed to develop the final assessment model.

A representative user experiment case is presented below:

- **Experiment dimension:** interaction level
- **Experiment objective:** to construct varying levels of interaction response delay to verify users' tolerance to system response waiting time in multi-modal real-time conversational AI services
- **Experiment scenario:** entertainment & gaming such as chess companion. Users engage in a chess game while receiving real-time guidance from a multi-modal conversational AI. The chess board state is streamed to the AI system for real-time inference, while users interact with the AI through voice in a multi-turn dialogue.
- **Pre-experiment result:** confirms the feasibility of the overall methodology. Small-scale pre-experiment results demonstrate a trend of declining user experience with increasing interaction delay, as shown in the figure below.

a) As interaction response delay increases, user experience shows a nonlinear downward trend. Within approximately 1 second, the decline in experience is relatively rapid, and after exceeding 1 second, the decline tends to slow down.

b) When interaction response delay exceeds approximately 2 seconds, user experience is significantly affected, and the experience score drops below the passing threshold of 3 points.

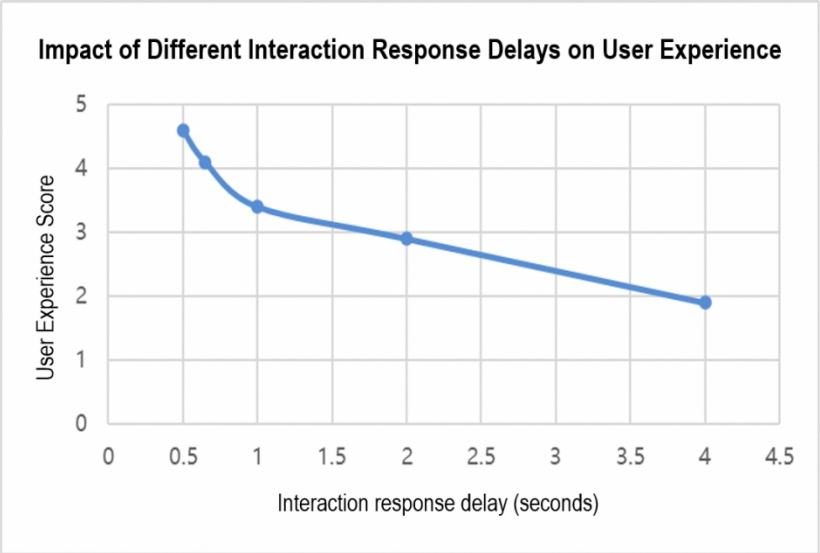


Figure 3-5 Figure 3-5 Interaction Scenario Test Result.

3.4 Mathematical Formulation of the Assessment Model

Balancing model completeness and modeling feasibility, a quantifiable, multi-dimensional QoE assessment model is proposed for AI multi-modal interactive applications. This model employs hierarchical mapping strategy to aggregate various influencing factors. The model framework is illustrated in the figure below.

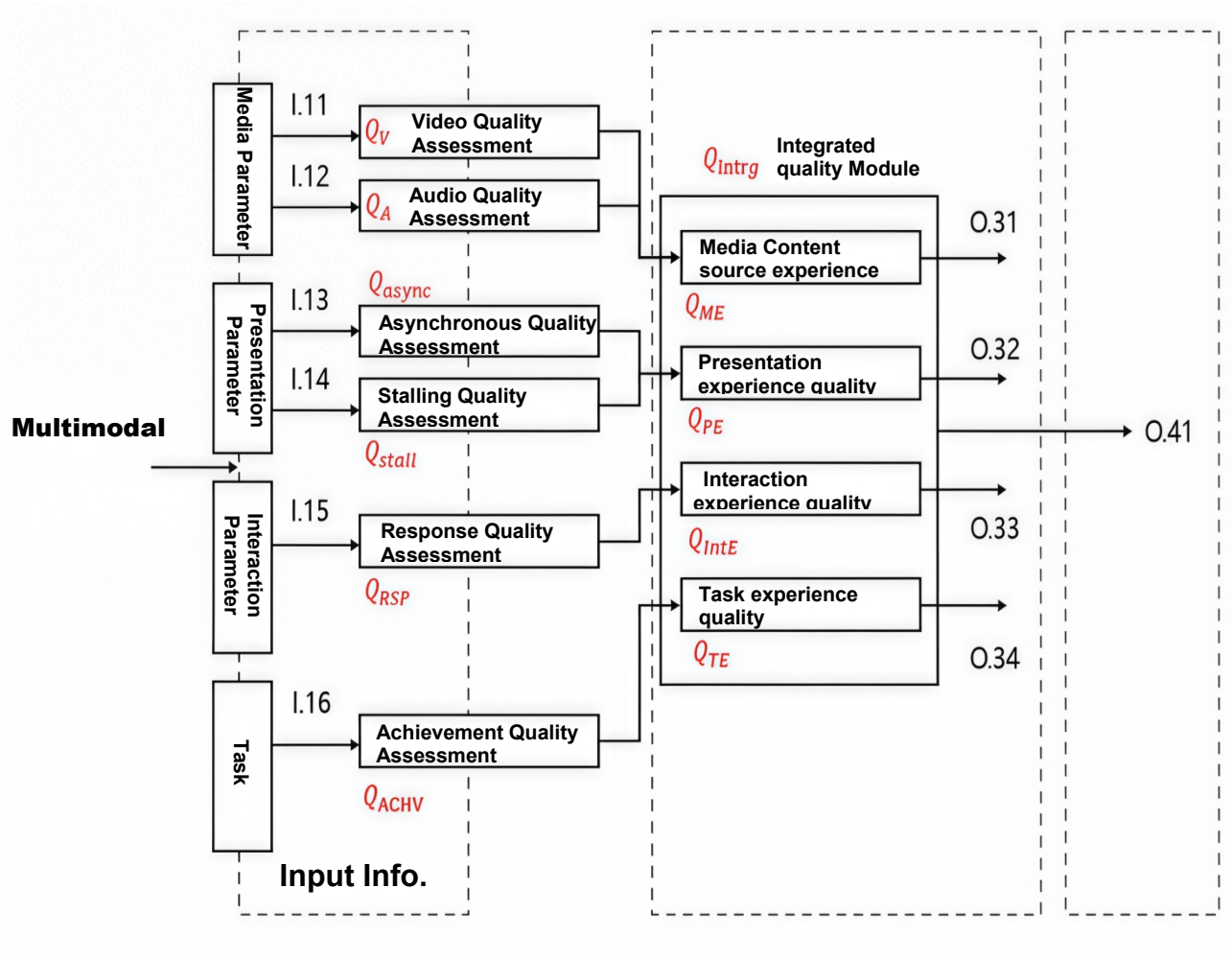


Figure 3-6 Evaluation Model Framework.

The multi-modal AI experience quality assessment model is defined as a function of four sub-dimensions:

- Content source experience quality (O.31), represented by Q_{ME}
- Presentation experience quality (O.32), represented by Q_{PE}
- Interaction experience quality (O.33), represented by Q_{IntE}
- Task experience quality (O.34), represented by Q_{TE}

The comprehensive AI experience quality (O.41) is calculated using the following formulas:

$$\text{AI experience quality} = f(Q_{ME}, Q_{PE}, Q_{IntE}, Q_{TE})$$

$$Q_{ME} = f(Q_V, Q_A)$$

$$Q_{PE} = f(Q_{async}, Q_{stall})$$

3.5 Application Value of the Assessment Model

The research work on this assessment model will bridge a gap in the industry regarding quantitative, end-to-end assessment of user experience for mobile AI applications, offering a scientific framework to guide industrial planning.

For mobile AI service providers and large model vendors, this experience assessment model enables continual optimization of service and product strategies based on empirical user data, fostering user loyalty and driving industry expansion. Crucially, as mobile AI applications rely on mobile networks for terminal-to-cloud connectivity, it necessitates advanced mobile network capabilities. By correlating user experience to network capabilities, operators can transition to a user-experience-centric network quality assessment methodology. This 'experience-driven' strategy for network construction, optimization, and operation is pivotal for ensuring the sustainable, high-quality evolution of future mobile networks.

3.6 Illustrative Service Profiles

To facilitate industry stakeholders in rapidly implementing this evaluation framework, standardized scenario profiles are summarized for four typical Mobile AI service forms. Each profile specifies the service characteristics, recommended deployment strategies, and core observation metrics, serving as a direct reference for operators, application developers, and equipment vendors to conduct experience assessment:

Profile	Typical use case	Likely placement	Representative metrics
Conversational voice AI	Care, FAQ, translation, search, personal assistant	On-device wake word; edge/cloud ASR and LLM	TTFT, E2E delay, interruption/barge-in handling, ASR quality, answer quality, task success
Mobile multimodal / vision assistant	Photo/video + voice/text guidance for field work, health, repair, or disaster response	Device capture; edge preferred for low latency; cloud for complex reasoning	Upload completion, frame integrity, multimodal sync, grounding quality, weak-uplink fallback, battery impact
Enterprise tool-using agent	Retrieve records, create ticket, call APIs, draft or execute approved action	Enterprise backend, private edge, secure cloud, or hybrid	Tool-call success, authorization success, grounded answer rate, recovery time, rollback/reversibility, audit trail
Background thinking agent	Long-form research, report generation, planning, summarization	Cloud or hybrid orchestration; cached context where useful	Task completion, progress visibility, continuity, cost/energy per successful task, failure recovery, user acceptance

Chapter 4:

Service Experience Assessment and Network Capability Mapping

4.1 E2E Service Capability Matrix

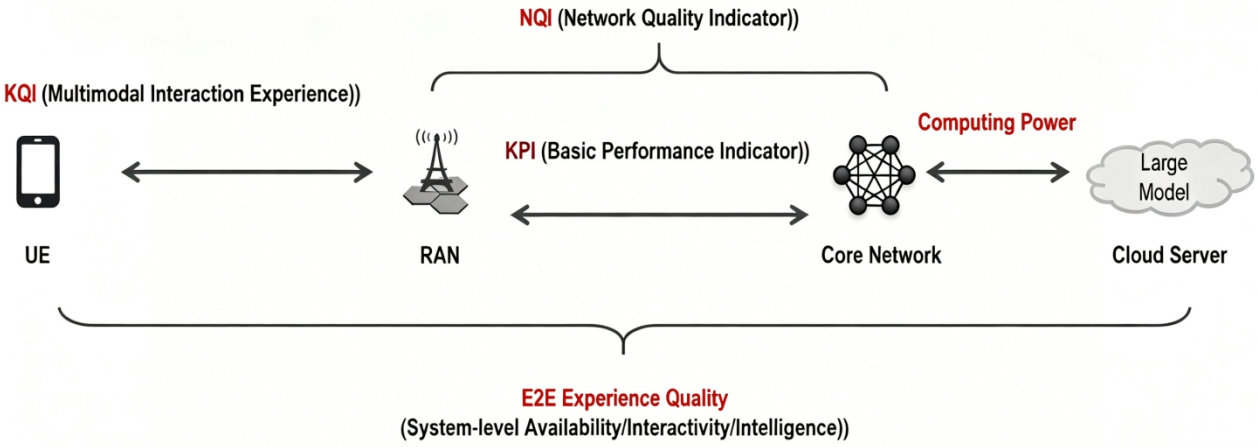


Figure 4-1 E2E Service Capability Matrix

Connection Metric Impact

In end-to-end service scenarios, connection quality is the cornerstone of user experience, with latency, throughput, and reliability constituting its core elements.

- **Latency and throughput:** Latency directly constrains the real-time performance of interactive services. Specifically, the transmission of AI token streams requires ultra-low round-trip time to ensure user-perceived fluency. Throughput determines the data transmission efficiency of multi-modal services, and insufficient bandwidth inevitably leads to stalling or loading failures.
- **Reliability and scheduling:** For packet-loss-sensitive services such as voice and video, high reliability is paramount for data integrity. However, traditional scheduling algorithms often struggle to accommodate the bursty traffic patterns inherent in AI token streams. This necessitates dynamic resource allocation mechanisms based on reinforcement learning to adapt resources to traffic patterns in real time, thereby mitigating the impact of traffic bursts.

To overcome connection bottlenecks, the industry has proposed multi-dimensional optimization solutions:

- **Architecture level:** Edge computing is introduced to offload computing tasks to edge nodes closer to users, physically shortening transmission paths and reducing core network latency.
- **Protocol level:** Adaptive modulation and coding (AMC) and hybrid automatic repeat request (HARQ) mechanisms are used to dynamically balance the transmission rate and reliability under complex channel conditions.
- **Service level:** Network slicing technology is leveraged to deliver customized connection services, ensuring that critical services receive priority resource guarantees.

Impact of Computing Power Metrics

The spatiotemporal distribution and scale of computing power deployment represent another key factor that determines service experience.

Category	Metric	Definition	Unit	Description
Computing power metrics	GPU utilization	The percentage of time the GPU core remains active during the statistical period, calculated as active GPU time divided by total sampling time.	%	Indicates the core's workload level.
	GPU throughput	The number of valid floating-point operations completed by the GPU per unit of time.	TFLOPS	Indicates actual utilization of hardware computing power.
Memory metrics	GPU VRAM utilization	Allocated GPU VRAM divided by total GPU VRAM.	%	Indicates memory pressure. High usage approaching its capacity limit may lead to OOM errors or request failures.
	VRAM bandwidth utilization	The percentage of current VRAM bandwidth relative to its theoretical maximum.	%	Indicates internal GPU memory access pressure.
	KV-cache hit rate	The percentage of historical tokens successfully retrieved from the cache to the total number of tokens.	%	A higher hit rate indicates greater KV reuse, resulting in more efficient VRAM utilization and less computation.
Scheduling metrics	Inference queue length	The number of requests waiting to be processed by the GPU, representing the total number of pending requests.	req	A persistently growing queue indicates insufficient processing capacity and increasing queuing latency.
	Request queue time	The time from when a request arrives at the service to when GPU processing begins.	Millisecond	Indicates the degree of request backlog. Long queue times typically signal system bottlenecks.
	Batch size	The number of requests processed concurrently in each GPU inference call.	req	Dynamic batching improves throughput, but excessively large batches may increase latency. Both average and maximum batch sizes should be monitored.
Response capability metrics	TTFT	The time from sending a request to receiving the first generated token.	Millisecond	Indicates prefill-phase latency and is critical for streaming interactions.
	Time per output token (TPOT)	The average time required to generate each subsequent token.	Millisecond	Indicates decoding speed and directly impacts user experience.
	Token throughput	The total number of output tokens generated per unit of time.	tokens/s	Indicates model output efficiency.
	Request throughput	The number of inference requests completed per second.	req/s	Indicates system processing capability.
	Goodput	The number of requests or tokens per unit time that meet specific SLOs (for example, latency requirements).	req/s	Goodput more accurately reflects user experience and business value than raw throughput.

- **Note:** The preceding description uses the GPU as an example and applies equally to other AI accelerators, such as NPUs and TPUs.

Edge computing and the distribution of compute power toward edge nodes: As Mobile AI workloads grow rapidly, traditional centralized architectures struggle to meet low-latency requirements. Edge computing significantly improves local processing efficiency by offloading computation to edge nodes. For example, in video analysis scenarios, performing image recognition at the edge significantly reduces cloud load and shortens response time.

- **Capacity configuration and joint optimization:** Computing power should be dynamically scaled based on predicted service demand to avoid queuing latency caused by under-provisioning or resource waste from over-provisioning. Research shows that jointly optimizing computing power and network infrastructure improves overall resource utilization and reduces energy consumption.

Looking ahead, as large AI models introduce increasingly diverse and complex requirements, building an elastic and scalable compute architecture—and enabling deep integration between compute resources and network slicing—will be essential to ensuring consistent service performance.

4.2 Network-Side NQI Construction

Evolution from Network Pipe KPIs to Service Experience NQIs

Network Quality Index (NQI) and KPI are fundamental metrics for assessing network service quality. Traditional wireless network assessment has long relied on KPI-based frameworks that focus on underlying network parameters such as signal strength, coverage, and QoS. However, these network pipe indicators often fail to accurately capture users' real service experience.

In recent years, with the growing adoption of User Experience Management (UEM), wireless network assessment has been shifting from a performance-centric approach to an experience-centric one. The NQI framework has emerged as a central component of next-generation wireless network assessment. By translating abstract network performance into quantifiable experience metrics—such as video loading latency, stall rate, and voice MOS—the NQI framework enables a transition from network pipe KPIs to service experience NQIs. This provides a more intuitive and accurate representation of users' satisfaction with network services.

Core Design Directions of the Network NQI System

On the network side, developing a modern NQI system centers on three key design directions:

- **Uplink experience optimization:** With the rapid growth of multimodal services—such as HD video uploads and real-time interactive applications—networks face increasingly tight requirements on uplink bandwidth and latency. Enhancing uplink experience has become a critical foundation for supporting emerging service models.
- **Differentiated experience assurance:** By introducing network slicing and priority-based scheduling, networks can deliver tailored service guarantees for different service types. For example, dedicated slice resources can be provisioned for services with bursty traffic patterns, such as AI token streams, to ensure transmission quality and responsiveness.
- **Deterministic experience delivery:** This focuses on maintaining stable and predictable service quality even under congestion or high-load conditions. Achieving this requires enhanced resource reservation mechanisms and dynamic load balancing techniques to ensure deterministic performance for core services.

Definitions of Key Network-Side NQI Indicators

To accurately quantify the above experience, the network side defines the following core NQI indicators and their corresponding calculation logic:

- **Uplink frame transmission delay:** The time interval from when the first packet of an uplink multimodal image or video frame is ready for transmission to when the last packet of the frame is successfully transmitted.
- **Downlink packet traversal delay:** The time interval from when the PDCP layer at the base station receives a downlink packet to when the packet is successfully scheduled for transmission at the MAC layer.
- **Multimodal frame size:** The total data size from the first packet to the last packet of an uplink multimodal image or video frame.
- **Multimodal frame rate:** The number of multimodal image or video frames transmitted per second in the uplink.
- **Multimodal frame interaction delay:** The combined uplink frame transmission delay and downlink packet traversal delay, reflecting the network's contribution to first-token interaction latency.
- **Effective uplink user-perceived rate:** A core indicator that reflects the user's actual uplink experience. It is calculated as: Effective uplink user-perceived rate = Multimodal frame size/Uplink frame transmission delay.

4.3 E2E Experience Assurance Enabled by Device-Network Coordination

To deliver an optimal experience for Mobile AI multimodal interaction services, dedicated E2E assurance capabilities must be established. The technology evolution must enable flexible coordination and scheduling as the industry shifts from single-modal services (single-stream services) to multimodal services (a single service that involves multiple concurrent streams). This ensures that Mobile AI can meet stringent requirements for multimodality, completeness, and real-time performance. This evolution is achieved through the following two key directions:

Direction 1: Accurate E2E Identification of Mobile AI Services and Differentiated, Fine-Grained Experience Assurance

Traditional QoS scheduling, which operates at second-level granularity, cannot meet the instantaneous high-concurrency demands of AI interactions. To address this, service and network domains must jointly introduce a new metric: short-term throughput.

- **Frame-level integrity assurance:** By jointly considering ultra-low latency on the order of 100 ms and instantaneous rate, the base station can achieve near frame-level control granularity.
- **Differentiated scheduling:** The network can dynamically adjust resource priorities based on the activation state of AI services, ensuring that multimodal data streams do not interfere with one another during bursty transmissions. This forms the foundation for real-time interaction.

Direction 2: E2E Identification of In-Service Data Types and Fine-Grained QoS Flow Coordination

Mobile AI services involve diverse data flows—including text, voice, images, files, and video—requiring the network to provide fine-grained flow-level awareness and differentiated handling.

- **New device-network coordination mechanism:** A newly co-defined media access control-control element (MAC **CE**) is defined.
- **Frame-level awareness and precise scheduling:** Devices proactively report key frame information through this mechanism, enabling the RAN to detect frame structures in real time.
- **Precise boundary control: The base station can accurately identify frame headers and tails to perform frame-integrity scheduling. Differentiated QoS policies are applied to different data types (for example, real-time sensitivity for voice, integrity for images), eliminating semantic gaps caused by packet loss or out-of-order delivery.**

This QoS mechanism driven by device-network coordination significantly enhances the transmission quality of various data flows in Mobile AI services, delivering a smoother and more stable user experience.

Chapter 5:

Future Outlook and Path Planning

5.1 Technology Evolution Path: From Current Assessment to 6G Standardization Inputs

The development of Mobile AI technologies exhibits clear stage-wise evolution. The progression of technologies and standards can be structured into three progressive phases:

Phase 1:
Optimization of the Current Assessment Framework and Token-Level Perception (5G/5G-A Phase)

The current Mobile AI assessment framework remains insufficient in areas such as multimodal interaction and quantitative assessment methodologies. Recent evolution efforts focus on refining measurement metrics and introducing NQI-based mappings. On the metric side, fine-grained indicators—such as TTFT and token smoothness—need to be incorporated into routine network optimization and SLA systems. On the network side, the RAN must acquire token-level perception capabilities, introducing new parameters such as Token Importance (TI) and Token Error Rate (TER). Leveraging the 5G-A air-interface packetization mechanism, the network can identify data boundaries and perform selective discarding or retransmission under degraded channel conditions. This ensures reliable delivery of high-value tokens and enables dynamic service control and enhanced transmission performance.

Phase 2:
Multimodal Coordination and Dynamic QoS Adaptation (Embodied AI and Edge Agent Phase)

As multimodal and multi-device coordination scenarios—such as embodied AI and edge agents—continue to mature, standards must evolve toward multi-stream coordination. Context shifts in AI services can trigger millisecond-scale fluctuations in QoS requirements, requiring the network to support dynamic service control. By frequently updating QoS parameters within air-interface protocol packets, or flexibly mapping a single QoS flow to multiple data radio bearers, the network can achieve priority-based scheduling of bursty traffic across text, voice, image, and other modalities. This enables deterministic user experience even in complex scenarios.

Phase 3:
Comprehensive Contributions to 6G Standardization (6G Intelligent Architecture Design Stage)

In the long term, Mobile AI experience requirements must be mapped to the core technical requirements of the 6G network architecture.

- **Endogenous intelligence:** 6G will enable adaptive, joint scheduling of network resources and AI tasks. Service–network coordination will be enhanced. On the data plane, services will expose data characteristics and transmission requirements (such as rate and latency) to the network. The network, integrated with the operating system, will provide multi-capability transport interfaces that allow services to invoke built-in AI engines and AI agents as needed, enabling automated configuration of network functions based on differentiated experience requirements.
- **Integrated governance**: Experience standards will evolve in alignment with the AI-native 6G architecture and cross-domain intelligent-agent coordination frameworks, ultimately forming a foundational component of integrated governance across AI applications, networks, compute power, and security within intelligent digital networks.

5.2 Industry Call to Action:Building a Unified Experience Benchmark and a Collaborative Mobile AI Ecosystem

Mobile AI is a complex, system-level undertaking that integrates communications, AI, semiconductors, and cloud computing. However, the absence and fragmentation of experience standards—despite Mobile AI application experience being the primary benchmark for assessing service quality—has become a critical bottleneck restricting large-scale, high-quality industry development. This white paper focuses on

the global challenge of standardizing Mobile AI application experience. From an industry-wide perspective, it identifies the need to build an open, inclusive, collaborative, and mutually beneficial Mobile AI ecosystem—grounded in joint standard development, guided by experience-first principles, supported by ecosystem collaboration, and oriented toward inclusive and sustainable growth.

- **Joint standard development:** Joint standard development is essential to preventing industry fragmentation. The global ecosystem should work together to define calculation methods and graded threshold levels for core experience indicators—such as TTFT, cross-modal asynchrony rate, and task success rate—based on shared industry interests. This will enable the creation of benchmark test suites and verification mechanisms that are mutually recognized across vendors and regions. At the same time, the industry should actively engage with international standards organizations such as 3GPP and ITU, incorporating innovations such as token-level perception and dynamic QoS adaptation into 5G-A evolution and 6G standardization. A dynamic standard-evolution mechanism should also be established to accommodate the rapid advancement of technologies such as multimodal interaction and embodied AI.
- **Experience-first:** Experience-first is the fundamental guiding principle for the development of Mobile AI. The entire industry chain should shift from network-indicator-driven development to experience-driven development, embedding user experience requirements throughout the full lifecycle of technology R&D, network deployment, product design, and service operations. Network operators should provide differentiated experience assurance—leveraging technologies such as network slicing and edge computing—based on quantitative mappings between service experience requirements and network capabilities. AI application developers and device vendors should optimize algorithms and product designs according to unified measurement benchmarks, while establishing continuous experience monitoring mechanisms to ensure that technological evolution remains aligned with real user needs.
- **Ecosystem collaboration:** Ecosystem collaboration is the key enabler for unlocking the full value of Mobile AI. It is essential to break down industry barriers and foster deep cooperation among communications operators, AI foundation-model providers, device vendors, and semiconductor companies. Together, these stakeholders should build an integrated device-network-cloud technical architecture and experience assurance framework, jointly define device-network coordination signaling mechanisms, and collaboratively advance the deployment and scheduling of edge-computing nodes. In parallel, the industry should establish open and shared technology platforms and test-and-verification environments to reduce R&D and testing costs, encourage cross-industry application innovation, and promote the deep integration of Mobile AI across vertical sectors such as manufacturing, healthcare, and education.
- **Inclusive development:** Inclusive development is an essential foundation for the sustainable growth of Mobile AI. All market participants—regardless of country, region, or enterprise size—should be enabled to participate in and benefit from industry development. This requires promoting open access to standards and technological achievements, lowering participation thresholds for small- and medium-sized enterprises (SMEs), and fostering a development environment that integrates enterprises of all sizes. Global technical exchange and cooperation should be strengthened to ensure that technologies and deployment experience are made accessible to less-developed regions, supporting the balanced rollout of Mobile AI networks and services and advancing efforts to narrow the digital divide. Furthermore, data security, privacy protection, and ethical governance must be embedded into technology-innovation and standardization processes to ensure the healthy and sustainable development of the industry.

List of Abbreviations and Symbols

Abbreviation	Full English name
A2A	Agent to Agent
ACR	Absolute Category Rating
AI	Artificial Intelligence
AMC	Adaptive Modulation and Coding
BPP	Bits Per Pixel
BSS	Business Support System
E2E	End-to-End
EUM	Experience User Management
FPS	Frames Per Second
GPU	Graphics Processing Unit
HARQ	Hybrid Automatic Repeat Request
ISO	International Organization for Standardization
ITU	International Telecommunication Union
KV Cache	Key-Value Cache
KPI	Key Performance Indicator
KQI	Key Quality Indicator
LLM	Large Language Model
MAC CE	Media Access Control Control Element
MCP	Multi-Core Protocol
MOS	Mean Opinion Score
MWC	Mobile World Congress
NIST	National Institute of Standards and Technology
NPU	Neural Processing Unit
NQI	Network Quality Indicator
NTN	Non-Terrestrial Network
OOM	Out of Memory
PDCP	Packet Data Convergence Protocol
QoE	Quality of Experience
QoS	Quality of Service
RAG	Retrieval-Augmented Generation
RAN	Radio Access Network

RSRP	Reference Signal Received Power
SA	Standalone
SINR	Signal to Interference plus Noise Ratio
SLA	Service Level Agreement
SLO	Service Level Objective
SR	Success Rate
TER	Token Error Rate
TFLOPS	Tera Floating-point Operations Per Second
TPOT	Time Per Output Token
TPU	Tensor Processing Unit
TTFT	Time To First Token
W3C	World Wide Web Consortium
5G-A	5G-Advanced
6G	6th Generation

Acknowledgements

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

Contributing Organization

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, Nokia, CICT Mobile, UNISOC, Rohde & Schwarz, Omdia, Noware, Showmac, Droi, Veezhen Consulting.

Supporting Organization

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, Honor, Keysight, KuaiShangYun, Leju Robotics, NXCLOUD, OPPO, Tongxin Micro.