GTI Metrics and Test Methods Towards AI device White Paper

http://gtigroup.org/

GTI Metrics and Test Methods Towards AI device White Paper

Version:	v_1.0.0
Deliverable Type	□ **Procedural Document** ☑ **Working Document**
Confidential Level	☑ **Open to GTI Operator Members** ☑**Open to GTI Partners** □ **Open to Public**
Program	5G Technology and Product
Project	Technology Evolution
Task	5G-A Technology
Source members	China Mobile, Qualcomm, Xiaomi, MTK, OPPO, vivo
Support members	Huawei, Unisoc, Keysight, R&S
Editor	Hang Ruan(CMCC), Jing Gao(CMCC), Jiachen Zhang (CMCC), Lei Cao (CMCC), Tianming Jiang(CMCC), Ting Wang (Qualcomm), Shuping Chen(Qualcomm), Yan Li(Qualcomm), Xiangkuo Xiao(Xiaomi), Daliang Sun(Xiaomi), Yingjun Li(Xiaomi), Mingyu Chen(MTK), Yuanyuan Zhang(MTK), Han Xiao(OPPO), Wendong Liu(OPPO), Sihao Shi(vivo),Bule Sun(vivo)
Last Edit Date	2024-11-23
Approval Date	

Confidentiality: This document may contain information that is confidential and access to this document is restricted to the persons listed in the Confidential Level. This document may not be used, disclosed or reproduced, in whole or in part, without the prior written authorization of GTI, and those so authorized may only use this document for the purpose consistent with the authorization. GTI disclaims any liability for the accuracy or completeness or timeliness of the information contained in this document. The information contained in this document may be subject to change without prior notice.

Document History

Date	Meeting #	Version #	Revision Contents
2024.11.25	GTI#41	V1.0	Initial Version

Table of Contents

Introduction

The new generation of information technologies represented by AI, big data, etc. are important industrial elements supporting the digital transformation of industries. The device market is accelerating to embrace the emerging trend of AI industry, and the development trend is good. AI mobile phones deploy AI models (such as GPT) on the end side to achieve multimodal man-machine interaction, presenting as non-single application intelligent mobile devices. Unlike the traditional approach of dispersing various intelligent functions on different apps in smartphones, AI smartphones integrate and link various functional applications in the form of AI agents through unified entrances such as intelligent assistants, thereby achieving user goals more efficiently. This design approach not only simplifies operations, but also provides users with a more natural and convenient multimodal man-machine interaction experience.

Driven by comprehensive upgrades on the model, chip, and operating system sides, AI smartphones are moving towards more efficient, intelligent, and personalized directions. On the model side, mobile phone manufacturers are actively exploring the application of large models, with a parameter count ranging from 1 billion to 13 billion. Since the second half of 2023, mobile phone manufacturers such as Xiaomi, OPPO, vivo, Honor, Huawei, Samsung, and Apple have successively applied large model capabilities to their products, and AI phones are accelerating their penetration. According to Counterpoint's prediction, the penetration rate of AI phones will increase from 11% in 2024 to 43% in 2027, which also means that mainstream models will gradually become AI phones in the future.

This report aims to promote the formation of an objective and unified functional and performance evaluation system in the industry for AI device, and provide evaluation scheme examples based on typical applications explored in some industries. Chapter 1 of this report provides an overview and definition of the AI device evaluation system, Chapter 2 introduces AI device evaluation solutions on Communication AI, Chapter 3 introduces AI device evaluation solutions on Generative AI, Chapter 4 introduce the development trends of AI device evaluation, and Chapter 5 is a summary and outlook of AI device evaluation.

This report is jointly written by China Mobile, Qualcomm, Xiaomi, MTK, OPPO, Vivo and other industry partners.

Executive Summary

The new generation of information technologies represented by AI, big data, etc. are important industrial elements supporting the digital transformation of industries, driving the AI device industry into a new era of ubiquitous intelligence and collaborative intelligence. Driven by technological upgrades and market demand, AI device is developing rapidly, but there are still certain shortcomings, such as the lack of AI device evaluation methods. This report aims to promote the formation of objective and unified AI device evaluation indicators for the industry, propose an evaluation system for the intelligence capability of 5G-A devices, and provide evaluation scheme examples based on typical applications explored in some industries. We hope to use the driving effect of "promoting research through evaluation" to drive the exploration and continuous investment of the industry in the application of AI device, and provide reference and guidance for the industry in planning, designing, evaluating, and verifying 5G-A AI device related technologies, solutions, and products.

Abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
AMF	Access and Mobility Management Function
OMC	Operation and Maintenance Center
NF	Network Function
MAPE	Mean Absolute Percentage Error
SPI	Shallow packet inspection
PFCP	Packet Forwarding Control Protocol
TLS	Transport Layer Security
DTLS	Datagram Transport Layer Security
SNI	Server Name Indication
HTTP	Hypertext Transfer Protocol
HTTPS	hypertext transfer protocol secure
SSL	Secure Sockets Layer
AMC	Adaptive Modulation and Coding
MU-MIMO	Multi-User Multiple-Input Multiple-Output
PDCP	Packet Data Convergence Protocol
PDSCH	Physical Downlink Shared Channel
CQI	Channel Quality Indicator
RI	Rank Indicator
DOA	Direction Of Arrival
MCS	Modulation and Coding Scheme
RRC	Radio Resource Control
PRB	Physical Resource Block
MTTR	Mean Time To Repair

1 5G-A AI Device Evaluation System

AI device refers to the introduction of AI capabilities to enhance wireless performance and reduce power consumption of devices, and to assist networks in achieving cost reduction and efficiency improvement in wireless signal processing and wireless resource management through end-to-end collaboration; Generative AI and large model technology are applied on the device side and efficiently collaborate with the cloud to provide highly personalized content and services to enhance user experience.

1.1 Evaluation system

With the evolution of AI device applications, the traditional indicator driven evaluation system is difficult to meet the diverse 5G-A intelligent evaluation scenarios, and there is a problem of difficulty in implementing a single evaluation indicator design. Therefore, the evolution of intelligent evaluation system needs to be gradually carried out and improved scene by scene, which is difficult to achieve overnight.

Based on this, this white paper innovatively proposes the "1-4-1" evaluation system architecture, which is based on typical scenarios and drives the four pillars of the evaluation system: evaluation environments, evaluation tools, evaluation indicators, and test interfaces. Based on the above four aspects, extract common and standardized related work for standardization, further promoting the industry consensus of AI device. This architecture guides the industry to gradually improve the AI device evaluation system by defining a universal methodology and top-level design.

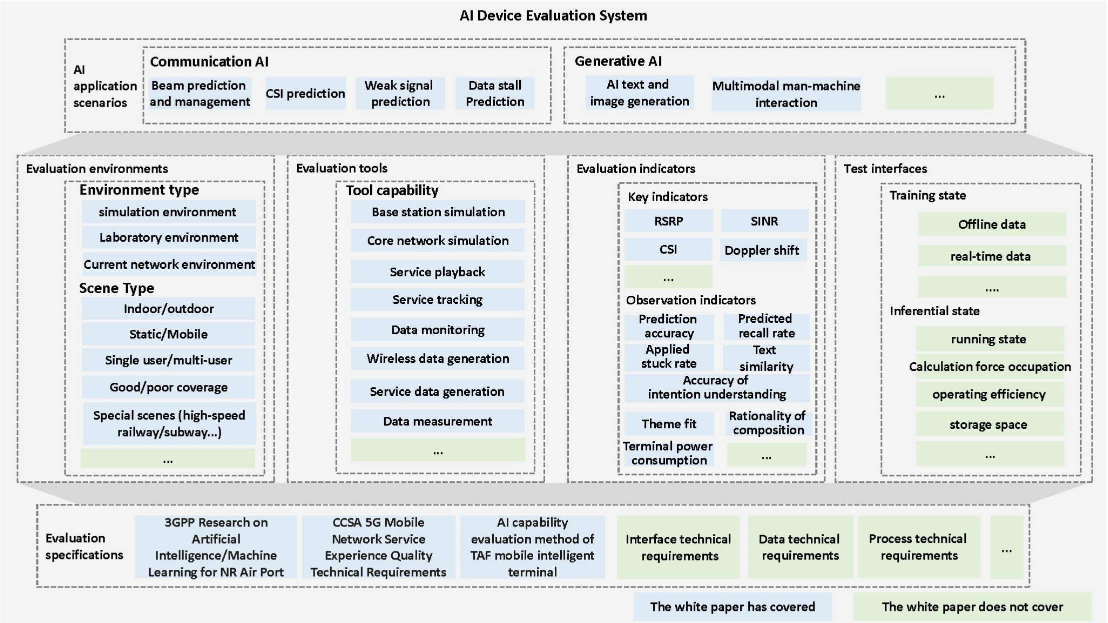


Figure 1 "1-4-1" Evaluation System Architecture

In the preceding figure, the blue sections represent content involved in the test cases of this white paper, while green areas denote recommended considerations for future evaluation work, though not explicitly detailed here.

1.1.1 Application Scenarios

AI device capabilities can be combined with actual scenarios to develop and design corresponding AI capabilities based on the characteristics of specific scenarios, further improving device capabilities. Based on this background, AI device evaluation needs to be combined with specific scenarios to construct evaluation examples one by one.

1.1.2 Evaluation Environments

The evaluation environment needs to be combined with the specific scenarios where AI capabilities play a role on the device side. For example, in the subway scene, predictive ability can be used to predict the time when users will enter weak signal areas in the future, and the device can make response strategies in advance. In the face of such endogenous intelligent solutions, testing and demonstration need to be conducted in the subway scene. AI has become a highly promising solution for complex scenarios due to its excellent fitting and generalization abilities. Therefore, in the process of evaluating AI devices, the selection and construction of the evaluation environment are particularly important. There are many wireless network environments in commercial scenarios, such as indoor and outdoor environments, high and low speed environments, interference environments, high-capacity environments, and so on. The evaluation environment should correspond one-to-one with the functional application scenarios of the device. In the process of constructing an evaluation environment, the applicable scenarios of the device should be fully considered, and appropriate environmental conditions should be selected to fully unleash the performance potential of the AI device.

1.1.3 Evaluation Tools

The evaluation tool needs to clarify the necessity and key role of the instruments used in the evaluation, and describe and require the functions of the tools. In response to the black box characteristics of AI, AI device testing instruments must have at least the following functions: core network function simulation, base station function simulation, service playback, service tracking, data monitoring, wireless data generation, service data generation, etc. When conducting evaluation work, most of the evaluation tools used in different intelligent scenarios are the same, such as some common data monitoring tools. However, for different scenarios, differentiated configuration of evaluation tools is still needed. For example, in service support scenarios, evaluation tools need to have the ability to input and replay service data; The end-to-end collaborative testing scenario requires capabilities such as core network simulation and base station simulation. For detailed usage, please refer to the typical scenarios in the following chapters.

1.1.4 Evaluation Indicators

The evaluation indicators need to clarify the key indicators and observation indicators in the evaluation. The key indicators are the internal support of the observation indicators, aiming to prove the effectiveness of intelligence through quantitative indicators. The traditional evaluation method based on a single evaluation index cannot cope with diverse AI device application scenarios. Therefore, this white paper follows the principle of "starting with the end and seeking common ground while reserving differences", and develops observable and diverse evaluation indicators based on the characteristics of different application scenarios to evaluate the degree of AI device. When selecting evaluation indicators, most commonly used indicators on the device side can be used, but they still need to be combined with specific scenarios.

The evaluation results of specific cases in subsequent chapters aim to demonstrate the feasibility of the method, with a practical verification process and no mandatory regulations in the industry.

1.1.5 Test Interfaces

The biggest feature of AI applications compared to traditional communication is black box, where key algorithms are not visible, because if you want to test the effectiveness of AI capabilities involved in a typical application, you need to present the relevant indicators in a standardized interface. This interface can collaborate with relevant testing tools to display the intelligence level of the AI capability in real time.

The testing interface is divided into offline output and real-time output. The training state related data includes offline data (such as model initialization parameters, training set data, etc.), real-time data (when the device capability supports it, the model can be trained and updated online based on some real-time device state data), etc. The data related to inference state includes operating status, computing power consumption, operating efficiency, storage space, etc. Based on this testing interface, timely updates of AI device models and real-time observation of AI model operation can be achieved in the future.

This work is still in the initial research stage, so it has not been mentioned in the specific scenarios of subsequent chapters. However, it should still be pointed out that the testing interface plays a significant positive role in opening up and standardizing related intelligent capabilities.

1.1.6 Evaluation Specifications

The evaluation standards need to clarify the correlation between the evaluation environment, evaluation tools, evaluation indicators, testing interfaces and industry norms and standards, and guide the scientific and standardized implementation of evaluation activities. At the same time, identify potential standardized evaluation content and improve the standardization construction of international and domestic evaluation systems. For example, device capabilities that are open in the form of interfaces need to be developed according to the corresponding technical system requirements in terms of interface definition, data types, input and output. The technical system can include data, interfaces, processes, and other related work to guide the standardized implementation of evaluation activities.

1.2 Tested object

For AI device, the tested objects are devices used for model inference, including smartphones, AI PCs, smart wearables, smart home devices, robots, drones, and other AI devices.

2 Communication AI Device Evaluation Solution

The application of AI device can be divided into two categories: the first is the application of communication AI, which enhances the wireless performance and reduces the power consumption of devices, and improves wireless performance and reduces costs through the collaborative assistance network of the device network; The second type is the application of end-to-end generative AI, which provides users with highly personalized content and services, enhancing the user experience. This chapter will present specific scenarios based on the "1-4-1" evaluation system architecture, including which specific solutions to adopt and how to conduct specific intelligent level analysis. Starting from methodology and top-level design, the industry will gradually improve the AI device evaluation system through specific scenarios.

2.1 AI based beam prediction and management

2.1.1 Application Scenario

Massive MIMO technology has been widely applied in 5G NR, which has designed a complete set of beam management processes, including beam measurement, reporting, and indication, as well as beam failure detection and recovery.

Beam management typically involves three processes: 1) finding matching base station transmit beams and device side receive beams through wide beam scanning; 2) Determine a more precise base station transmission beam through narrow beam scanning; 3) Determine a more precise device receiving beam through narrow beam scanning.

With the introduction of millimeter wave frequency, the number of beams gradually increases, and the increase in the number of beams will enhance the gain of beamforming. At the same time, the increased number of beams requires larger beam measurement reference signals and signaling overhead, as well as higher power consumption on the base station and device sides. Traditional closed modeling methods are very complex[1][2] and cannot effectively solve these problems. In recent years, the academic community has conducted in-depth research on using AI technology to achieve efficient beam management; Related research results indicate that AI technology can effectively reduce the reference signal and signaling overhead of beam management, and lower the power consumption of beam management for base stations and devices[3][4]. To this end, 3GPP studied the use of AI technology to improve beam management performance in Release 18, and has completed preliminary evaluations of performance gains and standard impacts, including beam reporting and beam prediction, in Release 18. In Release 19, standardization discussions are also underway in this direction, with the aim of using AI to reduce beam management overhead, minimize latency, and improve the accuracy of beam selection.

The AI based beam management methods mainly include spatial beam prediction and temporal beam prediction, which are used to reduce beam management overhead and delay. AI based beam management can be deployed on the network side or on the device side. The subsequent analysis is based on the introduction of AI beam management deployed on the device side.

Specifically, for AI based spatial beam prediction, the input of the AI model is the measurement results of some beams, and the output of the model is the probability of each beam becoming the optimal beam or the predicted beam quality among all candidate beams. Based on the model output, the optimal beam index or beam quality information can be directly determined among all candidate beams. Among them, the number of beams input by the model is less than the number of all candidate beams, and the width of the beams input by the model and the width of the candidate beams can be the same or different, depending on the specific application scenario, as shown in Figure 2 for example.

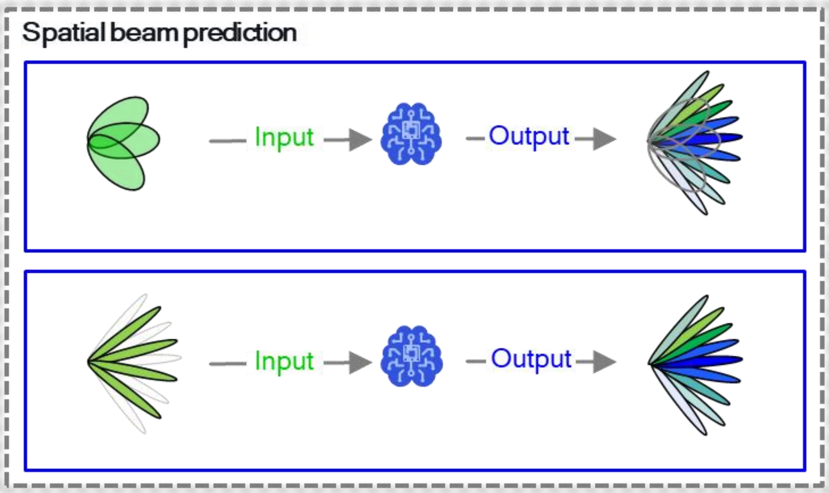


Figure 2 Case 1 Spatial Beam Prediction

In the 3GPP definition, AI based time-domain beam prediction is generally based on the RSRP values of beams measured at historical times, using past beam measurement results as inputs to the AI model to predict the optimal beam index or beam quality information for future times, as shown in Figure 3. Among them, the input of the model is the beam quality measured in the past time (i.e. the length of the observation window), and the output of the model predicts a time window that can be the same or different from the length of the observation window in the past time. In addition, the number of input beams and the number/width of candidate beams at the same time can be the same or different depending on the specific application scenario.

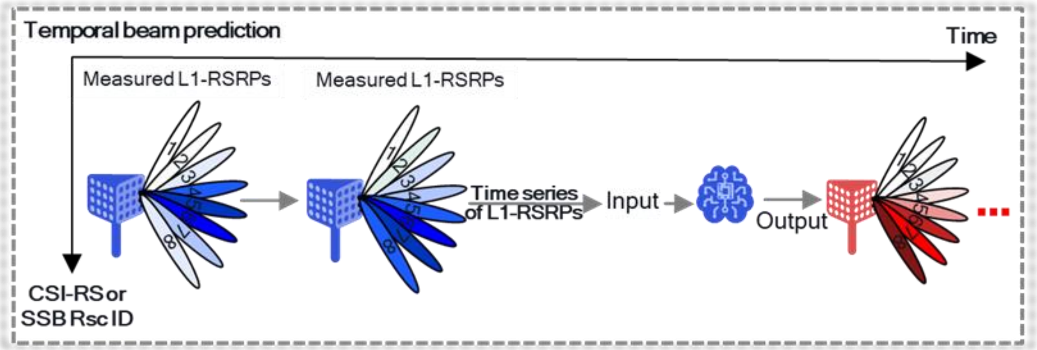


Figure 3 Case 2 Time Domain Beam Prediction

2.1.2 Evaluation Environment and Tool

At present, the AI based beam management testing method is still under discussion in 3GPP. It can refer to the 5G device MIMO performance testing, using the multi probe darkroom method. The testing environment consists of a base station simulator, a channel simulator, and a three-dimensional multi probe darkroom.

Arrange antenna arrays around the tested device in a 3D multi probe darkroom, and simulate the far-field environment originating from complex multipath by simulating the spatial distribution of arrival angles near the tested device. The signal is emitted from the base station simulator and propagates to the tested device through the multipath environment simulated by the channel simulator (i.e. spatial channel model). Before injecting all directional signals into the darkroom simultaneously through the darkroom probe array, appropriate channel losses such as Doppler frequency shift and channel fading are applied to each path. Then, the tested device antenna integrates the field distribution generated in the testing area and is processed by the receiver.

The testing tools include base station simulator, channel simulator, and multi probe darkroom. The base station simulator simulates 5G base stations to generate 5G network signals, while the channel simulator simulates multipath channel environments to generate various parameters set by the channel model, such as path loss, multipath fading, delay propagation, angle propagation, etc. The 3D multi probe darkroom is equipped with multiple transmitting and receiving antennas arranged according to certain rules to simulate different wireless channel environments, which can measure MIMO and radio resource management (RRM) performance under various fading conditions. The evaluation based on AI beam management will generate new requirements for evaluation tools, such as requiring more probes and channel simulators to generate more diverse multipath channels.

2.1.3 Evaluation Indicator

Currently, 3GPP is conducting in-depth research on the testing methods of AI for beam management in Release 19. Faced with this new feature, many methods that were originally applicable are facing challenges and new requirements. Currently, research is still in its very early stages, and more discussions are needed to find a balance between the complexity and accuracy of testing and reach consensus in the industry. The following analysis is still under discussion for reference.

After introducing AI models, the performance of beam prediction needs to be compared with traditional solutions. When conducting evaluations, it is necessary to use the existing beam pairing method to determine the beam pairs of the tested device and base station in different channel environments as the benchmark for performance comparison. Then, AI prediction is used to determine the beam pairs of the tested device and base station, compare them with the benchmark, and calculate the performance. On the other hand, it is also necessary to compare the cost of traditional solutions with the system after introducing AI, in order to evaluate the benefits of AI introduction in terms of system cost.

In the testing of beam management, there are two aspects that require significant changes:

On the one hand, the increase in the number of beams has put forward higher requirements for the testing environment and testing capabilities. In the testing discussion of 3GPP Release 19 beam management, a preliminary agreement was reached on the number of beams required for model testing. In the use case of beam management with the introduction of AI, the number of beams to be measured from the base station is 8 to 16, and the predicted number of beams to be measured from the device is 64 to 128.

On the other hand, the testing channel needs to be adjusted, and the transmission beam direction of the base station can also be dynamically adjusted. The previous measurement channels mainly used simple TDL channels[5], and AI models could easily predict the optimal beam pairing, but could not reflect the performance differences between different AI models.

Therefore, in order to verify the accuracy of beam quality prediction and reflect the differences in performance of different AI models, the beams in the test need to have sufficient randomness and spatiotemporal characteristics. Meanwhile, the prediction of airspace relies on observing beams with different arrival angles (AoA), and AI functional modules can learn spatial characteristics from these input data to infer the correlation between model input and output beams. Beam width is also a very important observation indicator in AI models used to infer the correlation between beams. Therefore, testing needs to consider the CDL channel[5], introducing different beam departure angles (AoD) and beam AoA, and the testing configuration also needs to include beam waveforms and beam gains with AoD as variables. An example of a simplified testing environment is shown in Figure 4.

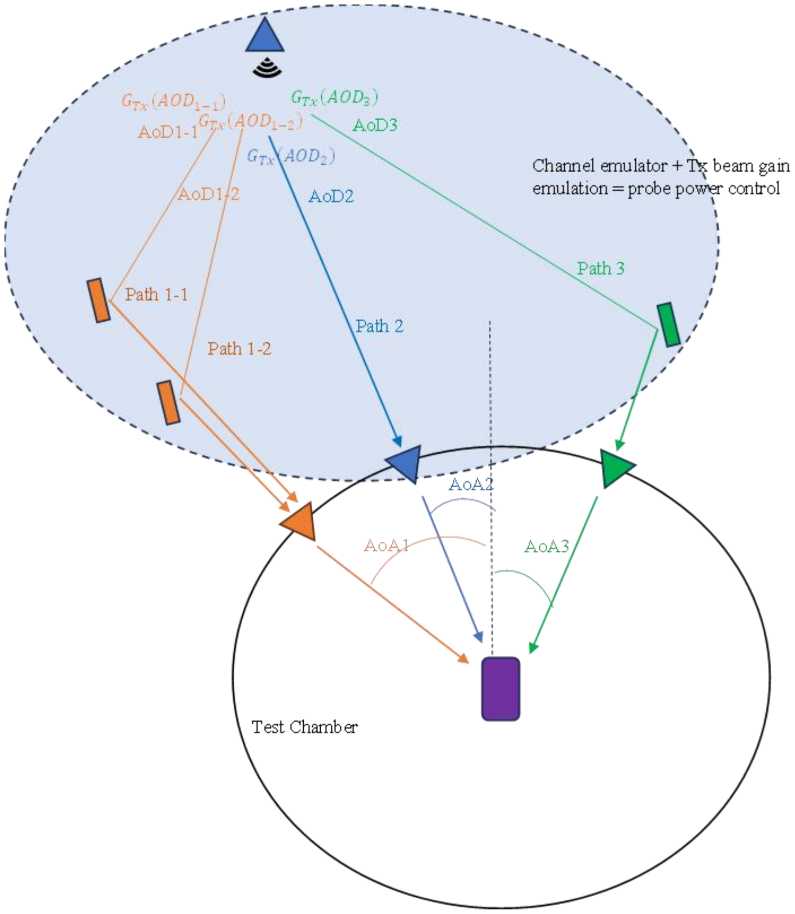


Figure 4 Example of Beam Management Test Environment

AI based beam prediction is evaluated based on prediction accuracy. Currently, there are four definition options, which will be discussed and defined by 3GPP in the future.

Option 1: RSRP accuracy

Option 2: Beam prediction accuracy

Top-1 (%): The percentage of "Top-1 strongest beam is Top-1 predicted beam"

Top-K/1 (%): The percentage of "Top-1 strongest beam is one of Top-K predicted beams"

Top-1/K (%): The percentage of "Top-1 predicted beam is one of the strongest beams in Top-K"

Option 3: The success rate of correct prediction, where the maximum RSRP among the first K predicted beams is greater than the RSRP of the strongest beam by x dB, can be determined by considering the relevant measurement accuracy

Option 4: Combination of the above options

2.1.4 Evaluation Specification

3GPP is currently researching testing methods for using AI for beam management in Release 19, and expects to complete specification development when the R19 standard is frozen.

2.2 AI based CSI prediction

2.2.1 Application Scenario

Obtaining reliable channel state information (CSI) is crucial in wireless communication systems. For example, MIMO systems require precoding or beamforming based on CSI at the base station to eliminate interference, and use CSI to select the optimal modulation and coding method in time-varying channels to improve throughput. However, in medium to high speed mobile environments, due to the relative movement of base stations and users, as well as the time-varying characteristics of the channel, CSI obtained through channel estimation and feedback often becomes outdated and cannot reflect the channel conditions at the scheduling time, seriously affecting the performance of wireless communication systems. Channel prediction is an effective technique to address the issue of CSI obsolescence. The device predicts the future CSI and reports it to the network side, and the network side schedules and pre encodes based on the predicted CSI.

Based on the temporal correlation of historical CSI, channel prediction can use past channel information to predict current channel information. The time-varying patterns of channels can be extracted from historical channel information through AI technology, and further channel prediction can be achieved. Specifically, as shown in Figure 5, the historical CSI is fed into the AI model, which analyzes the temporal variation characteristics of the channel and outputs future CSI.

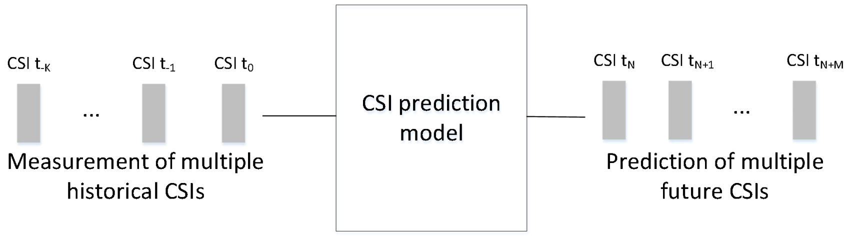


Figure 5 Schematic diagram of AI based CSI prediction

2.2.2 Evaluation Environment and Tool

The testing method for AI based CSI prediction is to first collect CSI data in the field, and then have the AI server perform data analysis.

Firstly, divide the actual outdoor area into several regions based on the location of the base station. The testing device access to a designated cell within an area and maintain synchronization status. The testing device parses the CSI parameter configuration of the cell through DCI information. To ensure the stability of data collection, the testing device moves along a fixed route and continuously collects CSI data from the community. Collect 3-5 sets of data (approximately 30 minutes) for each region, with corresponding time labels for each set of data.

After completing the field CSI data collection, perform subsequent analysis on the data in the AI server.

The testing of CSI prediction based on AI is completed by testing computers, testing devices and AI servers to collect and analyze data. During the testing process, the testing device is connected to an external testing computer, which controls the testing device to collect and store CSI data. The AI server is responsible for CSI data analysis.

2.2.3 Evaluation Indicator

Evaluation metric: The cosine similarity between the predicted CSI value at time+T output by the AI model and the actual measured CSI at time+T.

Firstly, collect a large amount of CSI data in the designated area and divide the data into training and testing sets. Train an AI model for CSI prediction using data of training sets and observe its performance on the testing sets. In order to demonstrate the advantages of AI based CSI prediction, the performance of non AI CSI prediction and no prediction scheme (sample and hold, S&H) based on autoregression (AR) algorithm was also tested on the same set of testing sets.

Figure 6 shows the SGCS gain of AI based CSI prediction compared to two non AI prediction schemes (S&H and AR). It can be seen that compared with S&H, AI based CSI prediction can achieve approximately 11% to 54% SGCS gain, and compared with AR prediction, it can achieve over 19% to 30% SGCS gain. Meanwhile, we found that the predictive performance based on AR is very poor in some cases, such as even worse than S&H in Cell 2. This is because there are channel estimation errors and phase discontinuities caused by RF links in actual channels, and AR is very sensitive to these damages.

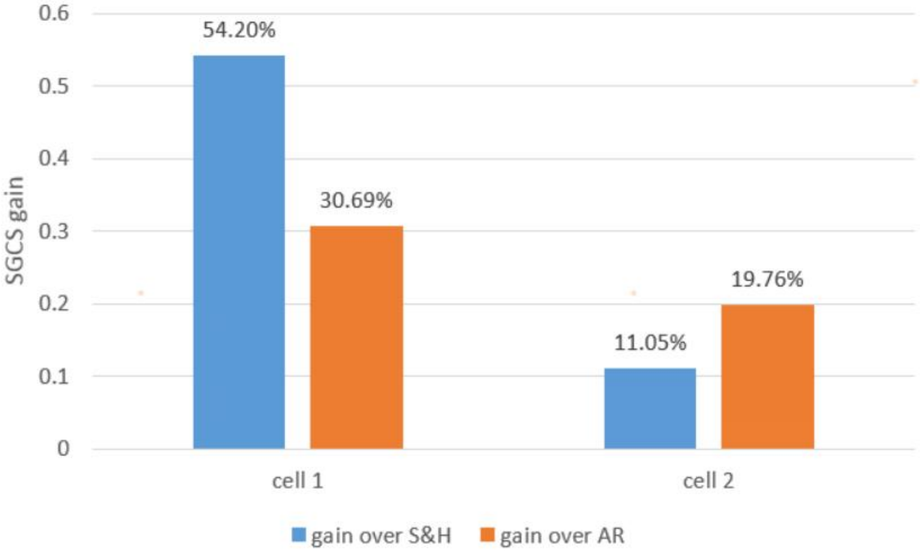


Figure 6 Comparison of Testing Performance for Channel Prediction

2.2.4 Evaluation Specification

This test case is mainly predicted by the device side for future CSI and reported to the network side. The CSI report will be applied to the interfaces in relevant standards. The evaluation of CSI prediction can refer to the cosine similarity described in the CSI feedback enhancement of 3GPP TR38.843 "Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface (Release 18)" [6]. This use case will undergo normal work in the latter half of 3GPP Release 19, and the testing method can also be developed based on the frozen version of 3GPP Release 19.

2.3 AI weak signal prediction

2.3.1 Application Scenario

In the modern mobile communication environment, users have extremely high requirements for the continuity and stability of data services, especially for the video viewing experience during the mobile process. One common challenge is that during commuting or travel, users may encounter areas with poor signal coverage, causing video playback to lag or be interrupted, seriously affecting the user experience. In order to solve this problem, AI based weak signal prediction technology for devices has emerged. Through machine learning, historical network signal parameters during user movement are collected as inputs for model inference and prediction of the user's future entry time into weak signal areas. devices can make response strategies in advance to improve the service experience of applications in weak networks.

In the field of AI prediction, various machine learning models can be used to process and analyze complex signal data to predict weak signal regions. Here are some commonly used machine learning models: Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Long Short Term Memory Network (LSTM), Linear Regression (LR), Random Forest (RF), M5 Decision Tree (M5T), Support Vector Machine (SVM).

Taking the subway scene as an example, using convolutional neural networks, the training of AI models requires a large amount of subway signal data. The model can learn the rules of signal changes and adjust its parameters to minimize prediction errors. The judgment logic of labels and the preprocessing of data are closely related to whether the model training converges. In addition, using techniques such as data augmentation, regularization, and dropout can prevent overfitting of the model and improve its generalization ability.

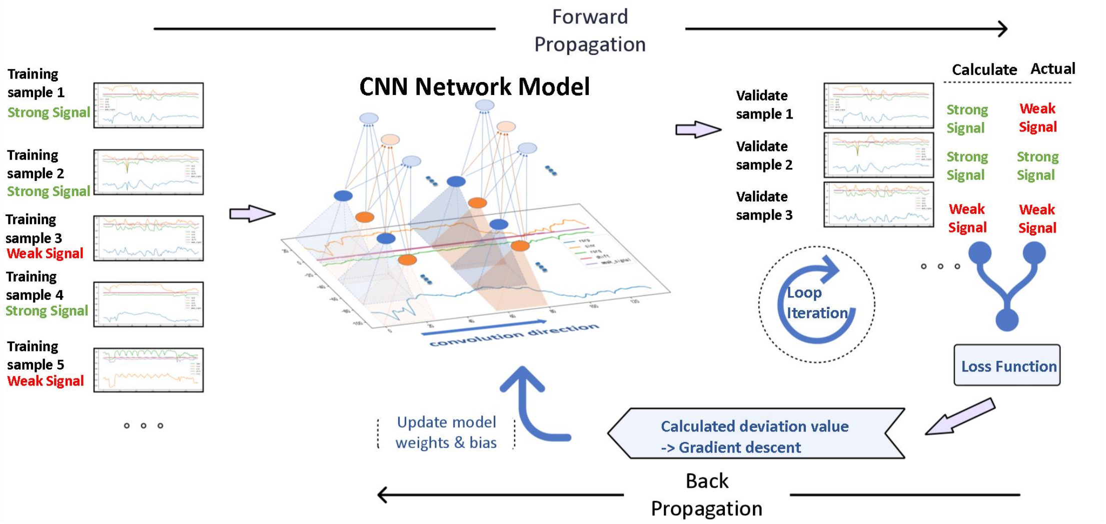


Figure 7 Training of AI Weak Signal Prediction Model

2.3.2 Evaluation Environment and Tool

In order to ensure the comprehensive and practical evaluation of the AI weak signal prediction system, tests were conducted on the current network and in the laboratory. Taking the subway scene as an example, the testing methods are as follows.

1. Field test: Focus on prediction accuracy and data stall optimization benefits, evaluate the predictive performance of the model by comparing it with actual signal quality. Firstly, deploy test devices on actual subway lines and install a third-party video app. Use road testing tools to track device behavior, strong and weak signal prediction results, real-time signal quality, and record short video playback using the system. During the testing process, the testers operated the third-party video app in different time periods and areas of the subway, continuously refreshing short videos for more than 20 minutes and flashing a new video every 5 seconds. After each AI weak signal prediction deduction, compare the signal prediction results of the device with the real-time signal quality, record the accuracy of the prediction, and calculate the prediction accuracy and recall rate.

2. Laboratory test: mainly evaluates data collection latency, model inference response time, system stability, computing resource consumption performance, power consumption

performance, and functional integrity. Firstly, using packet capture tools to capture data packets generated by users using third-party apps in the subway environment in the current network, these data packets contain key information such as signal strength, direction, and subway status. Then import these data packets into the testing environment of the laboratory, replay the data through an interface protocol tester, and simulate the network behavior and conditions of users using video apps in the subway.

The evaluation dimensions of the current Field testing cover model accuracy and optimized yield, and the required equipment only includes testing mobile phones, signal measurement and recording tools, and screen recording tools. All test indicators can be calculated by comparing the actual network signal strength, predicted signal strength results, pre optimized video playback and recording, and post optimized video playback and recording data.

Laboratory testing relies on replaying the signal data collected in the subway scene to simulate the real network scene of the subway, and can also add interference to the collected data to simulate weak signal scenes. The current network signal data can be collected through the installation of a road testing app on mobile phones. The road testing app can measure wireless diagnostic information on air interface and mobile application service quality (QoS) and quality of experience (QoE). After completing data collection, the current network signal data is played back using channel simulators and wireless platforms. At the same time, laboratory testing needs to consider power consumption testing, computing resource testing, as well as power meters and mobile system task monitoring apps to record the consumption performance during the model deduction process.

2.3.3 Evaluation Indicator

The typical evaluation indicators for AI weak signal prediction are prediction accuracy and prediction recall.

Prediction accuracy$Precision = \frac{TP}{TP+FP}$ includes:

- *TP* (True Positive): The number of correctly predicted positive categories.
- *FP* (False Positive): The number of incorrectly predicted positive categories.

Predicted recall rate$Recall = \frac{TP}{TP+FN},$ including:

- *FN* (False Negatives): The actual number of positive categories that were incorrectly predicted as negative categories

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F1 score ranges between 0 and 1, with higher values indicating better performance of the model

$$\text{Video preload revenue} = \frac{\text{compared machine stuck duration} - \text{test machine stuck duration}}{\text{compared machine stuck duration}}$$

2.3.4 Evaluation Specification

This scenario is mainly predicted by the user equipment to determine the signal strength at future times and reported to the application side. The interface reported to the application side belongs to the private protocol of the application and does not use the interfaces in relevant industry standards. At the same time, there are no relevant standards and specifications involved in the evaluation. If the base station and device jointly implement multi-user/single user scheduling in the future, it will involve the addition of interaction interfaces between the base station side and the device side.

2.4 AI data stall prediction

2.4.1 Application Scenario

With the rapid development and comprehensive coverage of 5G networks, the overall cellular network experience has been significantly improved. However, the network of cellular networks (high-speed rail/subway/highway, etc.) in high-speed mobile scenarios is complex and varied, and some core applications are sensitive to specific communication indicators, such as high bandwidth requirements for video live streaming scenarios and high latency requirements for online gaming scenarios. The device is conducive to AI/machine learning to detect data stall in advance. Before the data stall occurs, the user equipment can implement strategies such as cell selection, and also notify third-party applications to perform video preloading or reduce resolution. Using end-to-end AI to solve the pain points of users in high-frequency and high perception complex task scenarios, achieving a smoother user experience as they use it.

The device data stall prediction model can use current ten second signal parameters (RSRP/SINR/BLER, etc.) and full link quality parameters (TCP/IP and modem PCDP/RLC, etc.) to predict whether data stall will occur in the next ten seconds. Figure 8 shows the process of model generation: data processing, correlation analysis, sample processing, and model training. The model adopts the CNN neural network structure, which can extract deep features of the input and has excellent performance in classification training. To more accurately predict data stall results, the fixed trajectory full data stall prediction model will be subdivided into ten sub models, including travel scenarios (high-speed rail/subway/highway/general) and combinations of core parameters (bandwidth sensitive applications/delay sensitive applications/general). Based on sensor and network features, travel scenarios are identified, and communication indicators are distinguished by front-end app types. By combining travel scenarios and front-end application types, and combining end-to-end AI full chain data stall prediction models, high-frequency and high perception data stall is predicted in advance. The decision tree model is used to select the best end-to-end solution, effectively reducing data stall.

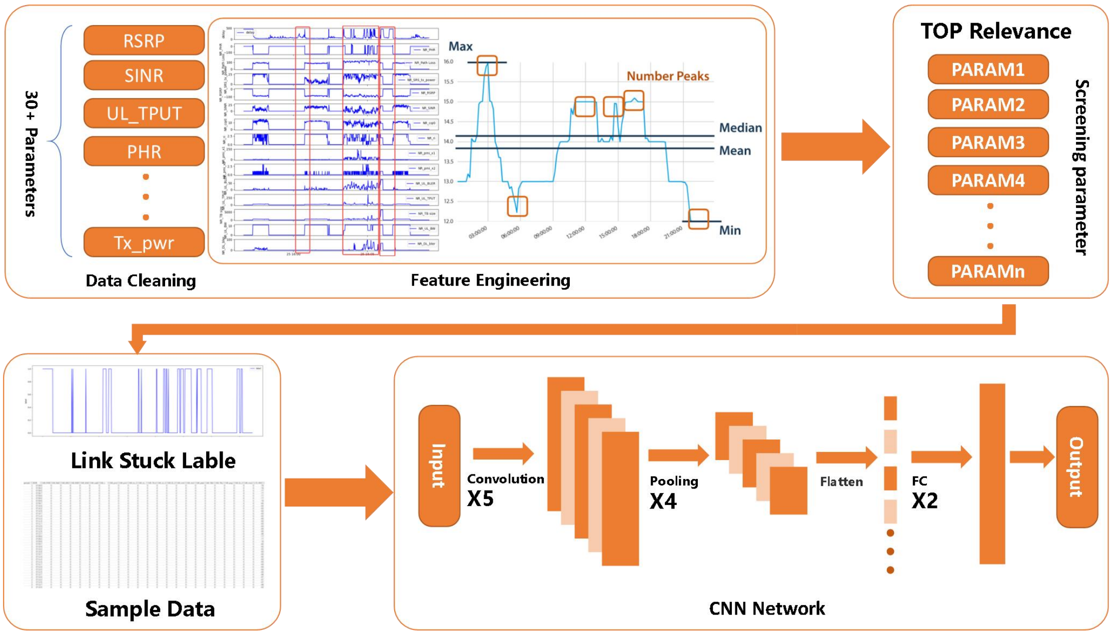


Figure 8　Schematic diagram of AI data stall prediction

2.4.2 Evaluation Environment and Tool

The test will be conducted on the current network. Firstly, a test device will be deployed on the actual fixed trajectory line and a third-party video APP will be installed. The device behavior will be tracked using road testing tools, and the system will record the playback status of the device. During the testing process, the testers operated the third-party video app at different time periods and areas on a fixed trajectory, continuously refreshing videos or WeChat video calls for more than 20 minutes, and refreshing a new video every 5 seconds. After each AI data stall prediction deduction, compare the data stall prediction results of the device with the real-time application data stall, record the accuracy of the prediction, and calculate the prediction accuracy and recall rate.

The current online testing and evaluation dimensions cover model accuracy and optimized profitability, requiring only two mobile phones, screen recording tools, and OCR data stall recognition scripts. All test indicators can be calculated through the data such as the stuck time and the stuck time during the use of third-party applications (such as Tiktok/WeChat video phone).

Laboratory testing relies on replaying the network characteristic parameters collected from the current network scene to simulate the real network scene of the subway. It is also possible to add interference to the collected data to simulate interference or weak signal scenes. The current network signal data can be collected through the installation of a road testing app on mobile phones. The road testing app can measure wireless diagnostic information on air interface and mobile application service quality (QoS) and quality of experience (QoE). After completing data collection, the current network signal data is played back using channel simulators and wireless platforms.

2.4.3 Evaluation Indicator

The typical evaluation indicators for AI data stall prediction are prediction accuracy and prediction recall rate.

Prediction accuracy $Precision = \frac{TP}{TP+FP}$ includes:

- TP (True Positive): The number of correctly predicted positive categories.
- FP (False Positive): The number of incorrectly predicted positive categories.

Predicted recall rate $Recall = \frac{TP}{TP+FN}$ includes:

- FN (False Negatives): The actual number of positive categories that were incorrectly predicted as negative categories

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The value of F1 score is between 0 and 1, with higher values indicating better performance of the model

Core application data stall rate reduction:

$$\text{Proportion of application data stall duration} = \frac{\text{Test machine data stall duration}}{\text{Test machine application duration}}$$

2.4.4 Evaluation Specification

Refer to Technical requirements for 5G mobile service experience quality developed by CCSA[7]. This standard defines the Key Quality Indicators (KQI) for 5G mobile network data service experience, which are applicable for evaluating the quality of 5G mobile network data service experience. The main content includes the definition of user level KQI for different types of services (including short video/long video, live streaming, instant messaging, web browsing, mobile gaming/cloud gaming, mobile payment, etc.).

3 Generative AI Device Evaluation Solution

3.1 AI text and image generation

3.1.1 Application Scenario

AI generated text refers to the use of AI technology to generate new textual content from existing datasets. It generates text fragments with certain logic and coherence through algorithms based on the input context, rules, grammar, and semantics. It can be used as a writing assistant, programming aid, story creation, news article generation, content creation, and other fields. It can also be applied to device large model applications in weak network or offline environments.

AI generated images refer to the use of AI technology to generate new images from existing datasets. At present, the task of image generation is divided into two types: text generated images and image generated images. The former generates images from text, takes natural language descriptions as input, and outputs images that match the description; The latter generates images from images and outputs stylized and special effects adjusted images based on the input images.

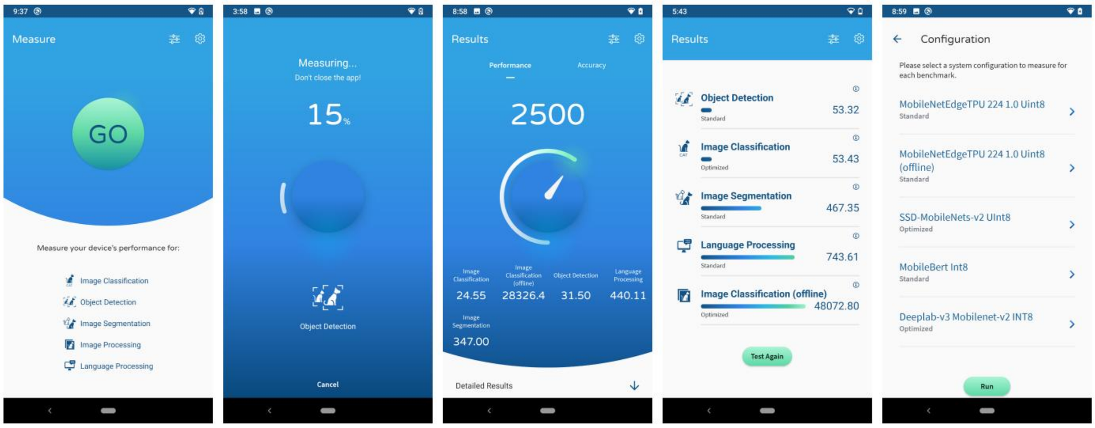
AI generated graphs are usually completed using a diffusion probability model. Generally, natural language description is used as input to output images that match the description. The commonly used text generation models currently include OpenAI's DALL-E 2, Google Brain's Imagen, and StabilityAI's Stable Diffusion.

AI generated images refer to the creation of new images based on existing ones. Stable Diffusion is one of the most widely used models in the field of graph generation. The "Miaoya Camera" incubated by Alibaba Entertainment in July 2023 became the first popular application of generative AI in China, with its underlying technology being Stable Diffusion. Graph generation technology can be used for:

- Image restoration, reconstructing lost or damaged parts from damaged or degraded images;
- Partial redrawing of the original image, such as changing faces or clothing based on the model's image;
- Painting assistance, such as drawing based on hand drawn sketches;
- Move objects, enlarge images, re compose and re edit photos (such as Samsung Picture Assistant)
- Modify the image style, such as changing the realistic style to anime style (already applied in vivo's Blue Heart model);
- Generate AI art photos based on selfies.

3.1.2 Evaluation Environment and Tool

For simple machine learning functions such as image classification, object detection, image segmentation, language understanding, etc., there are third-party evaluation software on the market such as Antutu, ML Perf, etc. that can perform mobile AI performance testing. The following figure shows an example of MLPerf testing[8].



(a) Startup screen (b) Running (c) Reporting results (d) Run details (e) Configuration settings

Figure 9 MLPerf Mobile app

However, the evaluation of generative AI such as large language models, especially on the device side, is still in its early stages. ML Commons introduced a test case for text generated image in MLPerf v4.1 version[9], initiating attempts at generative AI evaluation on the device side. In this use case, the testing of text generating image uses Stable Diffusion 1.5 as the model benchmark, the dataset uses MS-COCO 2014 captions, and the evaluation is conducted using a single query.

Further exploration is needed to evaluate more scenarios in the future.

3.1.3 Evaluation Indicator

The testing of AI text generation can be analyzed using a combination of objective and subjective indicators.

- Intention understanding accuracy

Use no less than 100 texts as input to generate text for the tested device. Determine whether the generated text can correctly understand the user's intention and text generation needs, and generate corresponding results. Calculate the accuracy of intent understanding according to the following formula.

$$W = \frac{S_1}{S} \times 100\% + 0.8 \times \frac{S_2}{S} \times 100\%$$

In the formula:

S - Number of testing services;

S1- Number of generated texts for precise understanding;

S2- Number of generated texts for fuzzy understanding.

- ROUGE-N

Use no less than 100 texts as input to generate text for the tested device. After splitting the results generated by the model and the standard results into n-grams, the recall rate is obtained, and Rouge-N is calculated to evaluate the similarity between the generated text and the reference text.

$$Rouge\_N = \frac{\sum_{S\in\{Reference\ Summaries\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Reference\ Summaries\}}\sum_{gram_n\in S} Count(gram_n)}$$

In the formula

N — n-gram, the number of bytes in the sliding window for text content, with a reference value of 2

$Count_{match}(gram_n)$ - The number of shared n-grams in the reference text and generated text

$Count(gram_n)$ - The number of n-grams in the reference text

In terms of subjective indicators, they can be analyzed separately through correctness, relevance, readability, logicality, and other aspects. Use no less than 100 pieces of text to generate text as input, such as editing documents, notes, emails, and messages, to enable the tested device to generate text. Check whether the generated text is related to the intention expressed in the text generation requirements through subjective MOS scoring. Using multiple person ratings, calculate the average MOS value after all tests are completed.

The application of AI generated images can also be evaluated by combining objective and subjective indicators. In objective indicators, FID is used to evaluate the quality of generated images, while CLIP is used to evaluate whether the content of generated images is consistent with the requirements[10].

- FID

Evaluate whether the semantic information of the generated image is related to the input text by measuring the feature vectors between the real image and the generated image through FID. The feature vector is composed of the output features before the classification layer of the reference distribution network. Assuming that the real distribution and the generated distribution are modeled as multidimensional Gaussian distributions, the FID calculation formula is

$$FID = \left\|\mu_r - \mu_g\right\|^2 + Tr(\varepsilon_r + \varepsilon_g - 2\cdot(\varepsilon_r\cdot\varepsilon_g)^{1/2})$$

In the formula

$(\mu_r，\varepsilon_r)$ - mean and variance of the true distribution,

$(\mu_g，\varepsilon_g)$ - mean and variance of the generated distribution

Tr - the trace of a matrix (sum of diagonal elements of the matrix)

- CLIP

Evaluate the quality of image generation by calculating the similarity of image features between the generated image and the reference image through CLIP.

$$CLIP = \omega\cdot\max\left(\cos\left(c, v\right), 0\right)$$

In the formula

c - Reference Photo Image Encoding in CLIP Model

v - Generating Image Encoding in CLIP Model

ω — Scalar coefficient

The subjective evaluation of AI generated images often uses MOS scoring, which can be conducted separately based on theme fit and composition rationality.

- Theme fit

Use no less than 100 different descriptions of text as test samples to generate corresponding images for the tested device. Check the degree of match between the generated image and the given textual description on the topic through subjective MOS scoring. Using multiple person ratings, calculate the average MOS value after all tests are completed. Individual ratings can be classified as follows.

Category 1: The image perfectly captures the theme described in the text without any deviation from the theme; The image conveys the emotions and atmosphere described in the text, and users can immediately recognize the theme and resonate with it

Category 2: The image effectively reflects the theme described in the text, with only a few details that may not fully align with the theme; The image is consistent with the textual description in terms of emotion and atmosphere, making it easier for users to identify the theme

Category 3: The image basically conforms to the theme described in the text, but there are some elements that are not very relevant or completely match the description; Images are acceptable in conveying emotions and atmosphere, but require users to think carefully before fully understanding the theme

Category 4: There is a significant deviation between the theme described in the image and the text, and some key elements do not match the description, which may cause misunderstandings; There are significant differences between images and textual descriptions in terms of emotions and atmosphere, making it difficult for users to identify the theme from the images

Category 5: The theme of the image and text description is completely mismatched, containing many elements that are irrelevant or incorrect to the description; Images cannot convey the emotions and atmosphere described in text, and users cannot recognize any information related to the theme from the images

- Reasonable composition

Use no less than 100 different descriptions of text as test samples to generate corresponding images for the tested device. Check the degree of match between the generated image and the given textual description on the topic through subjective MOS scoring. Using multiple person ratings, calculate the average MOS value after all tests are completed. The individual rating criteria can be classified as follows.

Category 1: The image composition is completely correct in spatial perspective, and the spatial relationships and depth perception of all elements are accurately presented

Category 2: The image composition is basically correct in spatial perspective, and the spatial relationships and depth of most elements are appropriately expressed. Only a few details need to be adjusted

Category 3: The image composition is generally correct in spatial perspective, but there may be some elements with inaccurate spatial relationship processing, which affects the overall coordination

Category 4: There are some errors in spatial perspective in image composition, resulting in inaccurate spatial relationships and depth perception of some elements, which affects the naturalness of vision

Category 5: There are serious errors in spatial perspective in image composition, and the spatial relationships and depth perception of most elements are chaotic, which seriously affects the overall visual effect of the image

3.1.4 Evaluation Specification

Generative AI is still in its early stages of rapid development. Previously, there were evaluation standards for intelligent applications such as machine learning, but evaluation standards for generative AI are still blank. At present, the CCSA organization TC11WG3 is discussing the research on the application and standardization requirements of generative AI and large models on the device side; The TAF organization WG7 is developing a group standard for the "Assessment Method for AI Capability of Mobile AI devices", which is discussing the evaluation method of generative AI and can be used as a reference.

3.2 AI multimodal man-machine interaction

3.2.1 Application Scenario

Generative AI brings more natural, comprehensive, and multi-dimensional man-machine interaction to devices, breaking the limitations of traditional single independent channel input methods and greatly enriching the dimensions of man-machine interaction. Multi modal understanding models allow users to communicate with devices using various data types such as text, images, sound, video, and sensors, improving decision-making accuracy and contextual understanding capabilities. At the same time, output methods can also become multidimensional and customized, providing users with a richer and more natural interactive experience. As a result, the way of man-machine interaction has undergone significant changes, shifting from the traditional "user instruction centered" approach to "user intention centered" approach. For example, when a user receives a message and copies it, the system will automatically analyze the semantics and extract key content, predict the next requirements and operations, and automatically go straight to applications such as memos and maps.

There are more and more discussions around AI multimodality, and multimodal models are gradually becoming more diverse. In August 2024, Microsoft launched the Phi-3.5 series models, which included multilingual and visual support. In August 2024, Google highlighted LMM (Large Multimodal Models) at its Made by Google event, which included the Gemini Nano model for multimodal input. In May, OpenAI launched its own multimodal model GPT-4 Omni. Previously, Meta also released the LLaVA (Large Language and Vision Assistant) model.

In addition to the support of multimodal models, smooth and sensory rich man-machine interaction also requires the assistance of other AI models. Take the conversation between users and virtual characters on mobile phones as an example for analysis. When users talk to the AI assistant, the speech is converted into text through OpenAI's automatic speech recognition (ASR) generative AI model Whisper. The AI assistant then uses the large language model Llama2-7B to generate text replies, and then uses the open-source TTS model to convert the text into speech. At the same time, the rendering of virtual avatars must be synchronized with voice output to achieve a sufficiently realistic user interaction interface. The device can use audio to create fused deformation animations to bring appropriate animation effects to mouth shapes and facial expressions. The smooth implementation of the entire process relies on the high computing performance and memory efficiency of the device hardware.

One of the key issues to be addressed when deploying large model/AIGC applications on devices such as mobile phones is how to achieve high-performance inference. Quantization is one of the most effective methods to improve computational performance and memory efficiency, and using low integer precision is crucial for energy-efficient inference. Multiple research works have found that for generative AI, due to memory limitations, Transformer based large language models often achieve significant efficiency advantages after quantization to 8-bit (INT8) or 4-bit (INT4) weights. Specifically, research has shown that many generative AI models can be quantified to INT4 models through quantitative research such as Quantitative Perception Training (QAT). Without compromising accuracy and performance, the INT4 model can save more power, achieving 90% performance and 60% energy efficiency improvement compared to INT8. In addition, in the process of deploying AI models to hardware architectures, compilers are the key to ensuring their efficient operation with the highest performance and lowest power consumption. Compiling includes steps such as partitioning, mapping, sorting, and scheduling of computational graphs.

When deploying a large model on the device, corresponding simplifications will be made to the model to ensure the efficiency of inference, so it is very important to evaluate the performance of the model application. Under the premise of considering hardware performance, the granularity and richness of evaluation content need to reflect the differences in performance of different AI models, and provide positive guidance for industrial development.

3.2.2 Evaluation Environment and Tool

When deploying a large model on the device, corresponding simplifications will be made to the model to ensure the efficiency of inference, so it is very important to evaluate the performance of the model application. Considering hardware performance, the granularity and richness of the evaluation content need to reflect the differences in performance of different AI models.

Multimodal man-machine interaction is a rapidly developing application, and methods for evaluating its performance are also being explored. The testing of multimodal man-machine interaction requires testing the combination of multiple AI tasks, while considering the use of sensors (such as location, behavior recognition status) and other information to evaluate various output capabilities, such as ringing, voice reminders, screen flickering, vibration, etc. Within the acceptable range of testing complexity, the evaluation model can effectively distinguish the AI capabilities of different devices. The comprehensive testing methods for these abilities are still in the early stages of research in the industry.

This article is based on the evaluation capability of a single AI task, considering testing multiple capabilities that the model can support separately, for future reference in multimodal model evaluation, including the evaluation of text to text, man-machine interaction, memory ability, etc.

Different from testing communication functions and hardware performance, there are some commercial AI model evaluation tools on the market for testing AI applications, and some mobile phone manufacturers have developed their own AI device evaluation tools. But for multimodal applications, specialized evaluation tools still need to be developed. When evaluating objective indicators, the evaluation tool needs to be able to record and compare various inputs such as text, voice, video, sensors (such as location, behavior recognition status), and the ability to record and compare multiple output forms.

3.2.3 Evaluation Indicator

Multimodal man-machine interaction is a rapidly developing application, and methods for evaluating its performance are also being explored. The current discussion mainly focuses on testing individual AI tasks, usually with only a single input such as text or speech. The testing of multimodal man-machine interaction requires testing the combination of multiple AI tasks, while considering the use of sensors (such as location, behavior recognition status) and other information to evaluate various output capabilities, such as ringing, voice reminders, screen flickering, vibration, etc. Within the acceptable range of testing complexity, the evaluation model can effectively distinguish the AI capabilities of different devices. The comprehensive testing methods for these abilities are still in the early stages of research in the industry.

Therefore, the introduction in this chapter is still based on the evaluation ability of a single AI task, considering testing the various abilities that the model can support separately, for future reference in multimodal model evaluation. The following introduction includes assessments of man-machine interaction and memory ability. Performance can be measured separately from objective and subjective indicators. For example, the efficiency of man-machine interaction below can be objectively evaluated, but the accuracy of interaction can only be considered through subjective MOS scoring methods. Considering that the evaluation model for AI is still in the early stages of discussion, the methods introduced below are still part of the current discussion and are for reference only.

1. Evaluation indicators for man-machine interaction

- man-machine interaction efficiency

Human computer interaction aims to recognize user intent and improve interaction efficiency. Interactive testing can be conducted no less than 10 times, including high-frequency scenarios such as taxi hailing, navigation, shopping, and social sharing. The number of steps taken by users to complete service recognition should be recorded, and the improvement ratio between intentional man-machine interaction and traditional man-machine interaction should be compared. The efficiency improvement should be calculated according to the formula.

$$F=(M-N)/M * 100\%$$

In the formula:

F - human-machine interaction efficiency;

N - number of steps for intentional man-machine interaction recognition;

M - Number of traditional interaction steps.

- Accuracy of man-machine interaction

Trigger service recognition by selecting information such as voice, text, and images. Select voice, text, and image target samples no less than 10 times each, including high-frequency scene content such as taxi hailing, navigation, shopping, and social sharing. Check whether the identified service meets expectations through subjective MOS scoring. Using multiple person ratings, calculate the average MOS value after all tests are completed. The individual rating criteria are shown in Table 1.

Table 1 Subjective scoring criteria for man-machine interaction accuracy

score	five	four	three	two	one
basis	All target samples and service recognition meet expectations.	Most of the target samples and service recognition meet expectations.	The general target sample and service recognition meet expectations.	A small number of target samples, service recognition meets expectations.	A very small number of target samples, and service recognition meets expectations.

2. Assessment indicators for memory ability

Based on different interaction abilities, specific abilities can also be evaluated in a targeted manner. If the interaction can involve multiple rounds of dialogue and the device can use temporary contextual information, then the model has short-term memory. The accuracy and completeness of short-term memory queries can be evaluated, as shown in the following example.

- Accuracy of short-term memory retrieval

Evaluate whether the memory traces retrieved are accurate. Prepare no less than 100 short-term memory query requirements, with a total quantity of N, and test the dataset including questions and tasks in conversations; Engage in dialogue with the intelligent agent, submit short-term memory query requirements one by one, and record the agent's response; During the conversation, record the number n of information correctly recorded by the intelligent agent. Calculate the query accuracy of short-term memory according to the formula.

$$\mathrm{Accuracy} = \frac{\mathrm{n}}{\mathrm{N}} * 100\%$$

- Integrity of short-term memory query

Evaluate whether the retrieved memory traces are complete. Prepare no less than 100 short-term memory query requirements, with a total quantity of M, and test the dataset including questions and tasks in conversations; Engage in dialogue with the intelligent agent, submit short-term memory query requirements one by one, and record the agent's response; During the conversation, record the complete amount of information m recorded by the intelligent agent. Calculate the integrity of short-term memory queries according to the formula.

$$\text{Integrity} = \frac{\mathrm{m}}{\mathrm{M}} \times 100\%$$

If during the interaction with the device assistant, the assistant can store the information learned during the model training process for a long time, then the assistant has the function of long-term memory. Long term memory is divided into personal attributes, service preferences, app attributes, behavioral habits, consumption preferences, content preferences, etc. according to the type of memory. In the evaluation process, targeted assessment of long-term memory can be carried out, and currently the subjective evaluation of long-term memory is mainly based on MOS scoring. At the same time, during the interaction process, the model can also perform planning reasoning and call other device apps, which can be evaluated through subjective and objective indicators based on the functions it possesses.

The testing of man-machine interaction should not only test the overall interaction efficiency and accuracy, but also be divided according to functions, and the required characteristics of functions should be tested in detail to reflect the differences between different models. With the rapid evolution of AI models, whether the granularity and richness of evaluations can map the performance differences between models is a direction that the industry needs to explore together.

3.2.4 Evaluation Specification

Multimodal models are still in the early stages of rapid development, and evaluation standards for multimodal applications and multimodal man-machine interaction are still blank. At present, the TAF organization WG7 is developing a group standard for the "Assessment Method for AI Capability of Mobile AI devices", which discusses the evaluation methods for intelligent man-machine interaction and intelligent agents and can serve as a reference.

4 Development Trends

With the continuous development of AI device capabilities, the evaluation system for AI device is also evolving in the following four aspects:

Innovation in testing methods: AI device testing methods need to be constantly innovated. For example, agile development methods in software engineering can be referenced to achieve continuous integration and testing of AI device algorithms. Meanwhile, testing methods based on simulation and emulation can also be explored to reduce reliance on actual data.

The formulation of testing standards: With the gradual maturity of AI technology, the testing standards for AI device will also be gradually improved. This will provide a unified evaluation benchmark for testers, improving the comparability and credibility of test results. At the same time, the formulation of standards will also promote the healthy development of the AI device testing industry.

Interdisciplinary collaboration: AI device algorithms involve multiple disciplinary fields, such as computer science, mathematics, statistics, etc. The future intelligent testing of devices requires more interdisciplinary collaboration to integrate knowledge and technology from different fields and jointly address challenges.

The integration of ethics and law: In future AI device testing, ethical and legal factors will receive more attention. Testers not only need to focus on the technical performance of AI device algorithms, but also need to ensure that they comply with ethical and legal requirements to achieve sustainable development of the technology.

The evaluation of AI device is facing unprecedented challenges and opportunities. By continuously innovating testing methods, developing testing standards, strengthening interdisciplinary cooperation, and incorporating ethical and legal factors, we have reason to believe that future AI device testing will become more mature and reliable, providing a solid guarantee for the widespread application of AI device.

4.1 AI Agent evaluation

The ability of AI to complete tasks is becoming increasingly strong, evolving from simple instruction execution to higher-order intelligent agents that autonomously disassemble targets and complete tasks, namely AI agents. An AI agent is an intelligent entity that can perceive the environment, make decisions, and perform actions. It has autonomy and adaptability, and can rely on the abilities endowed by AI to complete specific tasks, continuously improving and perfecting itself in the process.

Evaluating the accuracy of AI agents and multimodal generative AI can be done through various methods and metrics. The following are some common evaluation methods and steps:

1. Model Accuracy Test

Use standardized datasets for testing to validate the accuracy of the model for comparison with other models.

Dataset: Select recognized benchmark datasets such as ImageNet, COCO, etc.

Indicator: Evaluate using accuracy and precision indicators.

2. Quantitative Evaluation

Use automated evaluation metrics to quantify the quality of generated results.

BLEU (Bilateral Evaluation Understudy): Used to evaluate the similarity between generated text and reference text, commonly used in machine translation and text generation tasks.

ROUGE (Recall Oriented Understudy for Gisting Evaluation): Used to evaluate the recall rate of generated text, commonly used in summary generation tasks.

CIDEr (Consensus based Image Description Evaluation): Used to evaluate the quality of image description generation tasks.

SSIM (Structural Similarity Index): Used to evaluate the similarity between generated images and reference images.

PSNR (Peak Signal to Noise Ratio): Used to evaluate image quality.

3. Qualitative Evaluation

By manually evaluating the quality of the generated results.

Manual scoring: Invite experts or users to rate the generated results, evaluating their relevance, fluency, and creativity.

User survey: Collect users' satisfaction and opinions on the generated results through a questionnaire survey.

4. Multimodal Consistency Evaluation

Evaluate the consistency and correlation of generated results between different modalities.

Cross modal consistency: Ensure consistency and relevance of information between different modalities such as generated text, images, videos, etc.

Situational understanding: Evaluate whether AI can understand and generate information that is relevant to the context and background.

5. Task Completion Evaluation

Evaluate the performance of the generated results in practical tasks.

Task success rate: Evaluate the success rate of generated results in specific tasks, such as the accuracy of answers in question answering systems or the accuracy of image descriptions.

6. Real time Performance Evaluation

Evaluate the performance of generative models in practical applications.

Inference Time: The time required to evaluate the generated results.

Tokens per Second: Evaluate the processing speed of the generative model.

Bandwidth Utilization: Evaluating the bandwidth usage of a generative model during runtime.

7. Sustainable Performance and Power Consumption Evaluation

Generative models have high requirements for computing power and memory on the device, and the overall system is in a highly loaded state. In multimodal application scenarios, it is necessary to run multiple models simultaneously, which poses a more severe challenge to devices. Therefore, the implementation of AI agents and multimodal generative AI on devices requires testing the impact of sustained performance and power consumption on battery and body heating.

Testing can be measured through an external power source or by calculating the difference in the electricity meter through software after completing a series of tasks. Through long-term testing, performance and power consumption are recorded separately for each time period (e.g. test duration of 30 minutes, recording results every 3 minutes for a total of 10 records), which can reflect the initial performance of the device computing device, as well as the performance and power consumption changes after the device temperature rises.

4.2 Zero overhead superimposed-pilot design and evaluation

In wireless communication systems, pilot signals are crucial for ensuring the reliability and effectiveness of communication system links. 5G systems have always relied on a series of pilot signals, such as demodulation reference signals (DMRS), channel state information reference signals (CSI-RS), sounding reference signals (SRS), and phase tracking reference signals (PTRS), to support various functions including channel estimation, resource scheduling, link adaptation, beam management, and more. This type of pilot is generally designed as a series of predefined patterns and sequences, which are orthogonally allocated with data symbols in terms of time and frequency resources, meaning that the pilot competes with the data for limited transmission resources. This undoubtedly brings significant pilot overhead, thereby reducing the spectral efficiency of data transmission and limiting the throughput performance of the system. Especially in the design of 6G, complex scenarios such as larger antenna arrays and adaptation to higher speed scenarios will demand higher requirements for pilot design and overhead, leading to further intensification of wireless resource competition between pilots and data. From now on, it is urgent to consider designing new transmission strategies for pilots and data to address the aforementioned issues and challenges.

- **The relationship between new pilot transmission and data transmission**

The ability expansion of endogenous AI in the next generation wireless communication system, based on AI based air interface enhancement, is expected to show great potential in improving system performance. Especially with the introduction of AI processing modules at the receiving end, the powerful nonlinear processing capability will relax the orthogonal constraints that have always been followed in the pilot design at the transmitting end, which also brings the possibility of re exploring the resource allocation relationship between pilots and data. For example, a new framework or method can be explored to allocate pilots and data in a non-orthogonal manner on transmission resources, in order to reduce or even completely eliminate the independent resource overhead of pilots and the competition for limited transmission resources.

Superimposed pilot is a type of technology with great potential for application that expands the resource allocation relationship between pilot and data. Its core idea is to superimpose pilot symbols and data symbols at the transmitting end to break through the performance limitations of traditional orthogonal transmission methods. By simultaneously transmitting pilots and data on the same time and frequency resources, spectrum resource sharing can be achieved, significantly improving spectrum utilization.

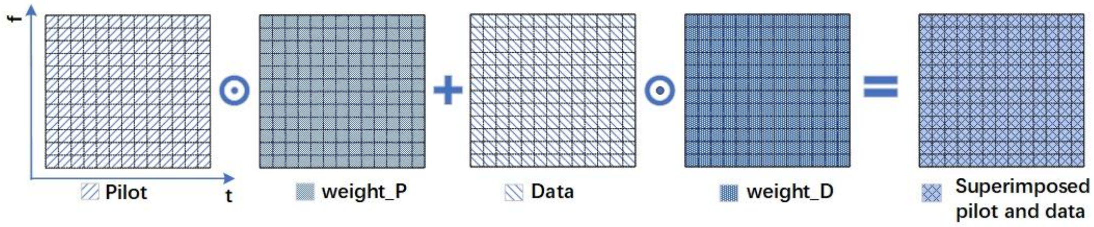


Figure 10 Non-orthogonal superimposed pilot construction method

As shown in Figure 10, at the transmitting end, consider weighting and summing the pilot symbol matrix and data symbol matrix within the time-frequency resources to obtain the superimposed symbol matrix for subsequent transmission. Obviously, in the above superimposed scheme, each transmission resource carries a portion of pilot information and data information, allowing the AI receiver to use the pilot information in the superimposed symbols to achieve channel estimation for each corresponding resource position, and also achieve high-frequency spectrum utilization data transmission on these resources. In addition, during the superimposed process, it is necessary to allocate the power allocation ratio between pilot symbols and data symbols reasonably, in order to ensure the final high-performance reception while balancing the data equivalent signal-to-noise ratio and channel estimation performance. At the receiving end, advanced AI receivers based on the superposition of pilot and data signals can be considered to achieve effective data reception. As shown in Figure 11, the input of the AI receiver is the received signal of the original information bit stream after modulation, non-orthogonal superimposed of pilots, and channel transmission.

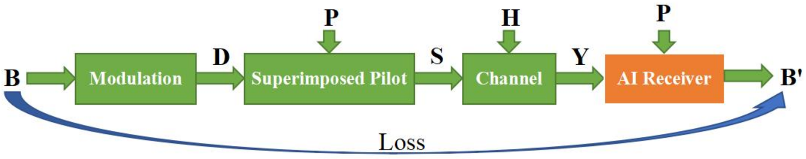


Figure 11 Non-orthogonal superimposed pilot transceiver system

- **The System Design Advantages of Superimposed Pilots for 6G**

The use of superimposed pilot schemes is expected to bring new changes and breakthroughs to the design of 6G systems:

Firstly, the introduction of non-orthogonal superimposed pilots reduces the overhead requirements of independent pilot resources and improves the spectral efficiency of the system. Although traditional orthogonal pilot schemes avoid mutual interference, they occupy a large amount of spectrum resources. With the further improvement of application requirements such as MIMO technology, high-speed scenarios, accurate channel estimation, and perceptual communication, the problem of pilot resource overhead still exists and will become increasingly apparent. The superimposed pilot design achieves the sharing of transmission resources between pilots and data by simultaneously transmitting them on the same time and frequency resources, greatly improving the utilization of limited resources. The independent resource overhead of the pilot signal in the time-frequency domain is zero, which means that more resources can be used for the transmission of data signals, thereby improving the throughput of the system.

Secondly, in traditional orthogonal pilots, the resource allocation pattern of pilot symbols is designed separately according to different transmission environments, which increases the complexity of system design and puts higher demands on network scheduling and management. The superimposed pilot scheme simplifies the allocation design of time-domain and frequency-domain pilot resources. In this non-orthogonal transmission method, since the pilot and data share the same resources, complex pattern design and resource allocation strategies are no longer required. Only how to optimize the power allocation of the pilot and data to ensure the performance of channel estimation and data transmission needs to be considered. This concise resource allocation design not only reduces the complexity of system implementation, but also improves the scalability and flexibility of the system.

In addition, as the superimposed pilot scheme distributes pilot and data information across all time-frequency domain resources, it can rely on the powerful nonlinear signal processing capability of AI receivers to perform more accurate implicit channel estimation and symbol detection based on the distribution of pilots and data across all resources. Whether in low-speed, high-speed, or more challenging ultra high speed mobile scenarios, superimposed pilots can flexibly adapt to different transmission requirements, have high robustness, and provide better system performance.

- **Testing and Verification of Superimposed Pilot Transmission and Analysis Scheme**

At present, the design and standardization of **superimposed** pilot signals still need to be carried out and refined. Due to the exploratory stage of the technical implementation plan and the instability of the testing plan, continuous follow-up research is needed.

The optional test cases are as follows:

- The test case parameters are shown in Table 2.
- The baseline scheme for comparison is the orthogonal pilot design and LMMSE receiver in the 5G system.
- Based on the test case parameter configuration shown in Table 2, complete the evaluation of the superimposed pilot transmission and parsing scheme.
- Compare the relative block error rate (BLER) performance of the superimposed pilot transmission and analysis scheme with the traditional baseline scheme under the given assumptions of mobile speed, transmission layer number, and modulation order, as well as the system throughput gain caused by avoiding independent pilot transmission, as the test evaluation results.

Table 2 Parameter Configuration of Superimposed Pilot Test Cases

parameter	to configure
Antenna configuration	1T1R, 32T4R
Number of transmission layers	1. 4
Channel model	UMa, CDL-C
Delay spread	300 ns
resource allocation	8RB, 52 RB

	12 symbols
carrier frequency	4 GHz
Subcarrier spacing	30 kHz
Modulation strategy	16/64/256/1024QAM
Encoding strategy	LDPC
Pre coding scheme	SVD
Moving speed	3/300/1200 km/h
Baseline Orthogonal Pilot Scheme	1 symbol, 4 symbols
Baseline receiver	LMMSE
Overlapping pilot power ratio	5%
Basic structure of the model	ResNet

4.3 Channel and Wireless Environment Semantic Communication and Evaluation

In recent years, semantic communication has attracted widespread attention from the academic community. In current semantic communication systems, the main focus is on extracting semantic information from data sources at the application level and transmitting it, such as images, text, videos, voice, etc. These semantic information can be extracted and utilized using artificial intelligence technology, thereby saving transmission resources and improving transmission performance during the transmission process. In the above process, the application layer source, semantic information extraction, and semantic space representation processes are not visible to the physical layer. The physical layer only focuses on the transmission process after resource mapping, and further considers cross layer optimization and the impact of new physical layer design, as shown in Figure 12 (a).

For wireless communication systems, especially for the internal design of the physical layer, the semantic communication between the physical layer channels of 6G systems and the wireless environment is a new problem and research direction. The design of semantic communication within wireless communication systems can consider using physical layer endogenous source information as input, as shown in Figure 12 (b). However, current research in this area is very limited, and the core issue is how to define and obtain physical layer source input for semantic information extraction.

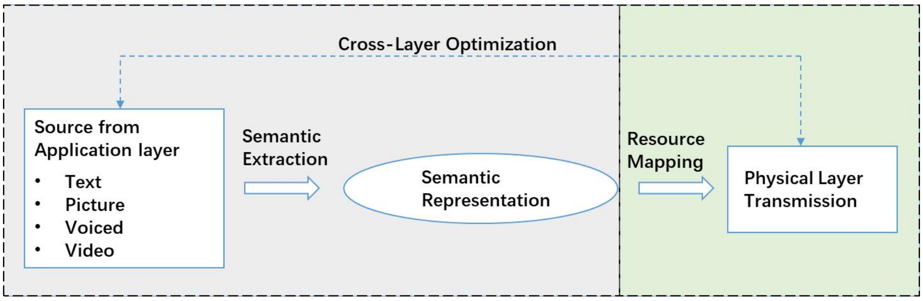


Figure 12 (a) Application layer semantic communication

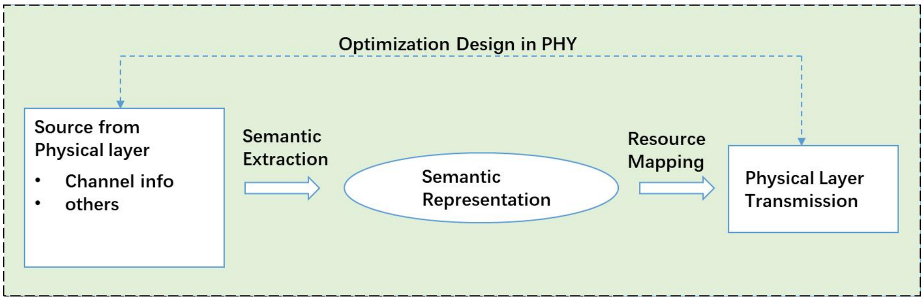

Optimization Design in PHY
Source from Physical layer
Channel info
others
Semantic Extraction
Semantic Representation
Resource Mapping
Physical Layer Transmission

Figure 12 (b) Physical Layer Semantic Communication

- **Concept and Implementation of Channel and Wireless Environment Semantic Information**

The semantic information of channels and wireless environments can be defined as the spatial/temporal/frequency domain characteristics of wireless channels and the distribution patterns of parameters such as power and phase brought by different wireless environments. The specific steps for obtaining and utilizing semantic information of channels and wireless environments include:

1) Base stations/devices obtain raw channel information by measuring different physical layer reference signals (such as CSI-RS);

2) Based on different purposes, different semantic information extraction methods (such as classical signal processing algorithms or AI methods) are used to transform the original channel into corresponding semantic spatial features;

3) Design corresponding transmission methods for physical layer semantic communication.

Channel state information is a typical semantic information of the channel and wireless environment. The device estimates the downlink CSI-RS channel and feeds back CSI to the base station for multi-user scheduling and downlink precoding. CSI contains multidimensional semantic information of wireless channels, which can be represented by traditional mathematical transformations to corresponding transformation domains for display or by AI model information extraction for implicit representation. For example, the semantic information contained in channel state information can include: Precoding Matrix Indicator (PMI) or channel feature vector that implies spatial frequency domain features, Channel Quality Indicator (CQI) that implies power distribution features, and Rank Indicator (RI) that implies the number of independent channels.

From the perspective of semantic communication, CSI contains semantic information of channels and wireless environments, mainly reported by users to the network side for downlink precoding and multi-user scheduling. The traditional CSI feedback process is shown in Figure 13. On the user transmitter side, it includes three processes:

1) Source compression, such as compressing the CSI feature vector calculated from user side measurements into a bitstream;

2) Channel coding, such as using polar codes or low-density parity check codes (LDPC) to complete channel coding;

3) Symbol modulation, for example, mapping the bit stream output after channel coding to constellation points through symbol modulation;

The modulated symbols are mapped to resources and then fed back to the uplink channel, which may contain non ideal factors such as noise and interference.

On the receiver side of the base station, there are three processes involved:

1) Symbol demodulation, for example, using a demodulation method corresponding to modulation to output the logarithmic likelihood ratio of the bits carried by the constellation points;

2) Channel decoding, for example, using a decoding method corresponding to channel coding to output an uncoded source information bitstream;

3) Source decoding, such as recovering the source information bitstream into CSI feature vectors.

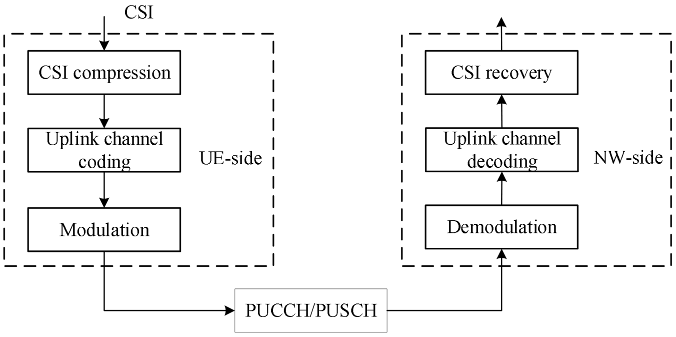


Figure 13 Traditional CSI Feedback Process

For traditional CSI feedback architectures, the CSI compression process is actually a lossy source encoding that introduces the loss of CSI source information, which cannot be recovered on the receiver side; After considering the fading and noise effects caused by the uplink channel, channel coding is used to reduce transmission errors and increase additional redundancy, thus occupying more uplink transmission resources; The loss caused by source information compression and the redundancy caused by channel coding are independently designed, and cannot achieve joint optimization in terms of transmission resources and performance.

Based on the joint source channel coding method, the CSI feedback architecture can be reconstructed as shown in Figure 14 (a). Deploy a joint encoder on the user side to achieve joint CSI compression and channel encoding functions. Its input can use channel feature vectors as source information, and the output is a joint encoded bitstream. Deploy a joint decoder on the base station side to achieve joint channel decoding and CSI recovery functions. Its input is the demodulated log likelihood ratio information, and its output is the recovered channel feature vector. Both the joint encoder and joint decoder can be implemented using AI models, and the training process adopts an end-to-end training approach. The potential advantage of this method is that the joint encoder and decoder can extract semantic information from CSI and learn CSI source distribution and uplink channel characteristics, achieving better feedback performance in balancing source information compression and channel coding to increase redundancy against noise.

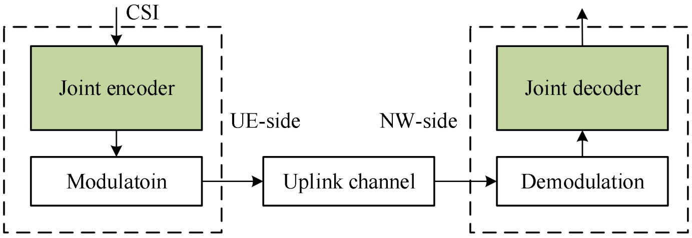


Figure 14 (a) Joint Source Channel Coding for CSI

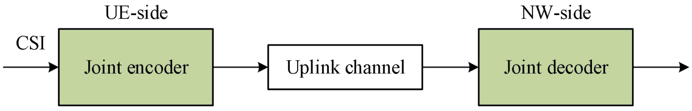


Figure 14 (b) Joint Source Channel Coding and Modulation for CSI

Furthermore, the CSI feedback method based on joint source channel coding can also include modulation and demodulation modules in the joint encoder and joint decoder. As shown in Figure 14 (b), the joint encoder implements CSI compression, channel coding, and modulation functions. Its input takes the channel feature vector as the source information, and its output is a complex sequence. Each complex value in the sequence is mapped to a physical resource for uplink transmission (equivalent to a modulated constellation point). Energy constraints can be applied to the complex sequence output by the joint encoder, such as constant modulus constraints or total power constraints. The joint decoder implements demodulation, channel decoding, and CSI recovery functions. Its input is the received symbols after channel equalization, and its output is the recovered channel feature vector. The potential advantage of this method is the end-to-end design of a through AI based on joint source channel coding and modulation, which can extract semantic information of the source and uplink channel characteristics, attempt to adaptively optimize the best matching method under the requirements of source coding, channel coding, and equivalent modulation order, and obtain better feedback performance.

- **Joint source channel coding testing and verification for semantic information of channels and wireless environments**

At present, the design and standardization of joint source channel coding for semantic information of channels and wireless environments still need to be carried out and refined. Due to the exploratory stage of technical implementation schemes and the instability of testing schemes, further research is needed.

The optional test cases are as follows:

Taking joint source channel coding for channel state information CSI as an example:

- The basic performance test parameter configuration is shown in Table 3, taking into account the actual uplink transmission link quality.
- The AI based CSI encoding scheme used in 3GPP R18 research serves as the baseline for comparison.
- Based on the parameter configuration of the test cases shown in Table 3, complete the evaluation of the CSI joint source channel coding scheme.
- Determine the normalized cosine similarity squared (SGCS) gain and corresponding system throughput gain of CSI joint source channel coding compared to traditional baseline schemes as the test evaluation results.

Table 3 Simulation Parameters of Joint Source Channel Coding for CSI

Public parameters		**to configure**
Downlink channel		UMa，32T4R
Upstream Channel		UMa，1T32R
CSI feedback layers		one
Model Architecture		Transformer baseline model
Upstream signal-to-noise ratio		(-20,3) dB
Number of uplink transmission resources occupied		64 Resource Particles (RE)
Baseline scheme parameters		**to configure**
-CSI encoder output -Bit rate -Modulation method	Baseline 1	32bit - 1/4 - QPSK
	Baseline 2	64bit - 1/2 - QPSK
	Baseline three	128bit - 1/2 - 16QAM
	Baseline 4	192bit - 1/4 - 64QAM
	Baseline 5	256bit - 2/3 - 64QAM
channel coding		LDPC without CRC
CSI Joint Source Channel Coding (JSCC Scheme 1)		**Parameter configuration**
Joint encoder output		128bit
modulation		QPSK

5 Summary and Prospects

As AI device continues to iterate, upgrade, and evolve, its application scope is constantly expanding, providing endless new space for the intelligent transformation of the entire industry and unleashing a continuous stream of new momentum. The widespread application of intelligence in various fields of social production and life will stimulate new experiences and bring earth shattering changes to humanity. The innovative application and industrial development of AI require the participation and collaboration of multiple parties, including industry, academia, research, and application. While grasping the trend of intelligent development in the industry, we should continuously pursue technological innovation, focus on engineering practice, and safeguard the benefits of intelligence for humanity. This white paper focuses on the architecture of the "1-4-1" intelligent evaluation system for 5G-A devices, evaluation solutions for high-value scenarios, and future evolution, providing guidance and theoretical basis for intelligent evaluation in the industry.

This white paper breaks away from the conventional indicator-based evaluation system and innovatively proposes a "1-4-1" AI device evaluation system architecture based on typical scenarios, promoting industry consensus and guiding the industry to gradually improve the AI device evaluation system.

6 References

[1] M. Giordani, M. Mezzavilla, C. N. Barati, S. Rangan, and M. Zorzi, “Comparative analysis of initial access techniques in 5G mmWave cellular networks,” in Proc. Annu. Conf. Inf. Sci. Syst. (CISS), Mar. 2016, pp. 268–273.

[2] V. Va, H. Vikalo, and R. W. Heath, “Beam tracking for mobile millimeter wave communication systems,” in Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP), Dec. 2016, pp. 743–747.

[3] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, “Efficient beam alignment in millimeter wave systems using contextual bandits,” in Proc. IEEE INFOCOM Conf. Comput. Commun., Apr. 2018, pp. 2393–2401.

[4] K. Satyanarayana, M. El-Hajjar, A. A. M. Mourad, and L. Hanzo,“Deep learning aided fingerprint-based beam alignment for mmWave vehicular communication,” IEEE Trans. Veh. Technol., vol. 68, no. 11, pp. 10858–10871, Nov. 2019.

[5] 3GPP TR 38.901, Study on channel model for frequencies from 0.5 to 100 GHz (Release 17)

[6] 3GPP TR 38.843 "Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface (Release 18)

[7] Technical requirements for 5G mobile service experience quality developed by CCSA

[8] MLPERF MOBILE INFERENCE BENCHMARK，https://arxiv.org/pdf/2012.02328v4

[9] https://mlcommons.org/benchmarks/inference-mobile/

[10] Improving Latent Diffusion Models for High-Resolution Image Synthesis (arxiv.org), https://ar5iv.labs.arxiv.org/html/2307.01952