世界互联网大会 World Internet Conference

中国电信 CHINA TELECOM | 研究院 CTRI

Outcomes of World Internet Conference
Think Tank Cooperation Program

Cognitive Alignment · Scenario Deepening · Ecological Collaboration: The Core Paradigm and Path of AI Evaluation in the Future

China Telecom Beijing Research Institute
China Telecom International Co., Ltd.

APRIL 2026


世界互联网大会
World Internet
Conference


中国电信
CHINA TELECOM
研究院
CTRI

Working Group

Chair

Yang Mingchuan, Big Data and Artificial Intelligence Research Institute,
China Telecom Beijing Research Institute

Vice Chairs

Wang Feng, Big Data and Artificial Intelligence Research Institute,
China Telecom Beijing Research Institute

Zhang Yuan, Big Data and Artificial Intelligence Research Institute,
China Telecom Beijing Research Institute

Lin Jianhui, Cloud Middleware Department,
China Telecom International Co., Ltd.

Members

Ding Peng, Big Data and Artificial Intelligence Research Institute,
China Telecom Beijing Research Institute

Zhao Jun, Big Data and Artificial Intelligence Research Institute,
China Telecom Beijing Research Institute

Liu Qian, Big Data and Artificial Intelligence Research Institute,
China Telecom Beijing Research Institute

Zheng Qiuhong, Big Data and Artificial Intelligence Research Institute,
China Telecom Beijing Research Institute

Wang Yuqiao, Big Data and Artificial Intelligence Research Institute,
China Telecom Beijing Research Institute

Zhao Yihan, Big Data and Artificial Intelligence Research Institute,
China Telecom Beijing Research Institute

Contact Email：zhengqh@chinatelecom.cn

Foreword

Against the background of the in-depth evolution of artificial intelligence toward generalization, scaling, and industrialization, AI evaluation has been upgraded from a single technical verification tool to core infrastructure that influences global technological competition, industrial layout, and governance rules. Based on a global perspective, combined with cutting-edge global theoretical innovations and practices, this paper proposes three core trends for future AI evaluation: the alignment of the essence of intelligence centered on "Cognitive Science", the in-depth penetration from general benchmarks to vertical scenarios, and diversified collaborative governance supported by platformization. This paper systematically analyzes the theoretical logic, global practical paths and core industrial values of each trend, and introduces typical global cases to provide forward-looking and operable think-tank references for global policymakers, research institutions and industries, so as to promote the development of AI evaluation in a more scientific, practical and governance-effective direction.

Contents

I、Global Strategic Positioning and Evolution Logic of AI Evaluation

(I) AI Evaluation: Definition and Connotation

AI evaluation is not an isolated form of assessment. Its core system has gradually expanded and evolved from early large model evaluation, and has now formed a comprehensive evaluation scope covering various AI forms such as large models, agents, AI application systems, and embodied intelligence. By definition, AI evaluation is a comprehensive activity that relies on scientific theoretical frameworks, standardized indicator systems and systematic technical methods to conduct quantitative assessment and qualitative analysis on core dimensions of various AI systems, including capability boundaries, performance, scenario adaptability and security risks. The value of AI evaluation is not limited to the well-known ranking lists. Its greater core value lies in integrating evaluation deeply into the whole process of AI R&D and production by building professional evaluation capabilities and developing standardized evaluation tools. It not only provides precise guidance for R&D optimization, but also builds a solid line of defense for security risk investigation. Ultimately, it offers objective and credible decision-making basis for industrial selection and regulatory governance of AI systems, serving as a key bridge connecting AI technology supply and industrial demand.

From the perspective of the classification system, the current mature AI evaluation system has formed a multi-dimensional division standard:According to evaluation objects, it can be divided into five categories: general large model evaluation, industrial large model evaluation, AI agent evaluation, multi-modal AI system evaluation, and embodied intelligent system evaluation.According to the evaluation life-cycle, it can be divided into three stages: pre-evaluation during R&D, compliance evaluation before launch, and continuous evaluation during operation.According to core evaluation dimensions, it can be divided into five directions: capability evaluation, security evaluation, compliance evaluation, efficiency evaluation, and fairness evaluation, forming a three-dimensional evaluation connotation covering all dimensions and the entire life-cycle.

From the perspective of connotation, the core value system of AI evaluation consists of three key components:Firstly, performance measurement at the technical level, focusing on general technical indicators such as accuracy, response speed, non-hallucination rate, and robustness of models and systems, which serves as the fundamental basis of the evaluation system. Secondly, value adaptation at the industrial level, emphasizing scenario-specific indicators including response accuracy, knowledge retrieval capability, and content generation quality of AI systems in vertical industries, realizing the deep integration between evaluation and industrial demands. Thirdly, risk prevention and control at the governance level, covering core requirements such as ideological alignment, privacy protection, and ethical compliance, which constitutes the bottom line for the safe and standardized development of AI technologies. With the evolution of artificial intelligence toward multi-form and full-scenario applications, the connotation of AI evaluation has expanded from single-dimensional performance verification to a comprehensive, full-chain evaluation system covering technology, industry, and governance.

(II) AI Evaluation: Core Hub for Global Competition and Governance

At present, artificial intelligence is in a critical development stage of accelerated iteration and all-domain penetration. Its evolution trend not only defines the value boundary of AI evaluation, but also drives the continuous upgrading of the evaluation system, which is

reflected in three core trends: Firstly, shifting from static testing for single tasks to dynamic adaptability evaluation integrated with cognitive science. As the parameter scale of large models keeps expanding and their capability boundaries continue to extend, they gradually gain cross-domain, cross-modal and multi-task general intelligence, creating an increasingly urgent demand for evaluating the essence of intelligence. Secondly, evolving from technical research and development to large-scale industrial application. With AI technology deeply integrated into thousands of industries including government affairs and manufacturing, scenario-based adaptability has become a core criterion for measuring technological value. Thirdly, moving from innovation-oriented development to equal emphasis on innovation and standardized governance. Countries around the world have successively issued AI governance policies, making security, trustworthiness, ethics and compliance a prerequisite for AI deployment, which strengthens the core role of AI evaluation in risk prevention and control.

The continuous evolution of AI technology and the constant expansion of application scenarios have laid the value foundation for the continuous iteration of AI evaluation. It has not only become an important pillar supporting technological innovation and industrial development, but also gradually grown into a core infrastructure underpinning global AI technological innovation, industrial implementation, and governance norms, emerging as an "invisible battlefield" for global technological competition and a "foundation for formulating" governance rules (as shown in Figure 1). Technically, it determines the direction and efficiency of AI R&D and guides the allocation of global innovation resources. Industrially, unified and credible evaluation standards are critical to breaking information barriers and reducing the cost of technology adoption. From the perspective of governance, the evaluation system acts as a core carrier for translating ethical principles disand security requirements into operable indicators, directly influencing the distribution of course power in global AI governance. From the EU AI Act, which takes compliance assessment as a prerequisite for high-risk AI systems, to various countries promoting localized AI governance frameworks with security and capability evaluation systems as core supporting measures, all these practices confirm its strategic importance.

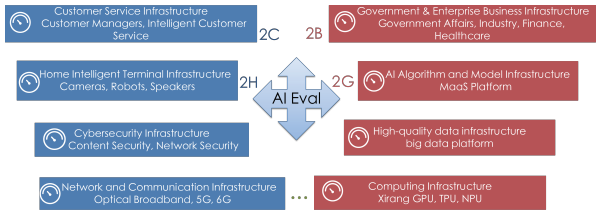

Figure 1: AI Evaluation Infrastructure

(III) The Evolution of AI Evaluation: From Technical Verification to Ecological Empowerment

Since the emergence of large models, the development of AI evaluation has gone through three stages: The first stage focuses on the single core performance of large models (such as language generation accuracy and knowledge QA correctness), addressing the basic question of whether AI is usable. The second stage shifts to the evaluation of general capabilities of large models (such as multi-modal understanding and complex reasoning), responding to the core demand of whether AI is good to use. Currently, it is entering the third stage, whose core is to solve the problem of how to scale up the deployment of large models safely, fairly and efficiently. The evaluation scope has expanded from technical performance to the essence of cognition, scenario adaptation, governance compliance and other dimensions, showing the distinctive features of deeper theory, finer scenarios and more ecological collaboration. [figure 2]

The first stage spans 2022–2023, representing the single- performance evaluation stage in the early days of large models. It focuses primarily on the basic performance evaluation of large language models. The landmark events include the release of GPT 3 (2022) and the rapid iteration of early large models. During this stage, large models had just emerged and their technologies were not yet mature. Evaluation mainly concentrated on basic performance metrics for single tasks, such as fluency in language generation, accuracy in knowledge question answering, recall rate in text classification, etc. The core goal was to answer the fundamental question: Can large models accomplish basic language tasks? The evaluation objects were mainly early general purpose large language models, and the evaluation method was dominated by offline testing on static datasets. As a whole, evaluation was positioned only as an auxiliary tool for technical verification of large models, without involving assessments of complex cognitive capabilities or scenario adaptability.

The second stage spans 2024–2025, marking the general capability evaluation stage during the iteration period of large models. It focuses on evaluating the general capabilities of large language models and multi-modal large models. The landmark events include the release of new generation large models such as OpenAI o1 series and DeepSeek R1, as well as the wide adoption and upgrading of general evaluation benchmarks including MMLU, AIME, and GPQA. During this stage, large models have gradually gained cross domain and multi-modal general processing capabilities. AI evaluation has shifted from single task performance testing to multi task general capability assessment, covering multiple dimensions such as language understanding, logical reasoning, multi-modal fusion, knowledge reserve, and contextual coherence. The core goal is to evaluate how strong the general capabilities of large models are, with ranking lists becoming the mainstream presentation form in this stage. However, with the rapid evolution of large model technologies, critical defects have gradually emerged in the evaluation system: a certain proportion of the world’s mainstream general evaluation benchmarks have been included in the training data of major large models, resulting in severely distorted test results. The contradiction between high benchmark scores and low practical application performance of large models has become increasingly prominent, driving the large model evaluation system into a brand-new evolutionary stage.

The third stage spans from 2025 to the present, representing the full-stack, full-life-cycle evaluation stage during the large-scale industrial deployment of large models. Its core feature is the comprehensive upgrading of AI evaluation from a single tool to an ecological infrastructure, which is also the core research stage of this paper. In this stage, large models have shifted from technological iteration to large-scale industrial application. AI evaluation has broken through the traditional positioning of "performance testing" and is fully carried out around the entire life cycle and application chain of large models, evolving toward three directions: alignment with the essence of cognition, deep exploration in vertical scenarios, and ecological collaborative governance. The evaluation objects have expanded from basic large models to a full chain including industry-specific fine-tuned large models, AI agents, multi-modal large model applications, and large-model-driven embodied intelligence. The evaluation cycle has been extended to the full life cycle covering R&D, fine-tuning, deployment, and operation. The value of evaluation has extended from technological verification to industrial empowerment and governance support, making it a core infrastructure for the large-scale and healthy development of large models. Here, "full-stack" mainly refers to covering the entire hierarchy of "basic large models –

industry fine-tuned models – large model applications". Together with "full-life-cycle", it forms a dual dimension of "space + time", improving the core connotation of the large model evaluation system.

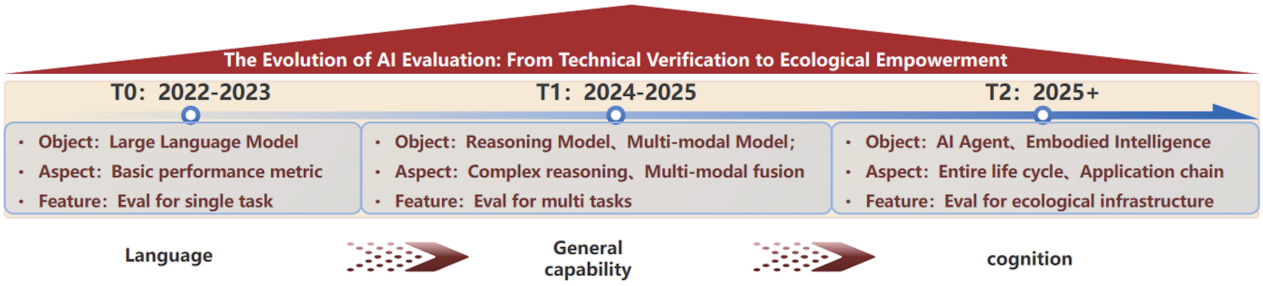

Figure 2: Development Stages of AI Evaluation

II、Trend 1: Cognitive Alignment –"Cognitive Science" Reconstructs the Theoretical Foundation of AI Evaluation

(I) Connotation of the Trend: The Fundamental Leap from "Measuring Performance" to "Measuring Intelligence"

Traditional AI evaluation takes task completion as the core indicator, which essentially measures the model's pattern matching ability rather than its genuine intelligence level. As large models evolve toward general intelligence, the "black box" problem and the limitations of capability assessment have become increasingly prominent: models may achieve high scores on benchmark tests but lack core cognitive abilities such as commonsense reasoning and causal judgment. Academician Li Deyi proposed that the four basic patterns of human cognition — experience pattern driven by memory, reasoning pattern driven by knowledge, creative pattern driven by association, and discovery pattern driven by hypothesis — form the foundational framework for cognitive formalization. This formalization provides theoretical support for the architecture of a new generation of artificial intelligence systems. By abstracting human cognitive patterns into computable structures, cognitive formalization enables machines to simulate processes including experience accumulation, logical reasoning, creative association, and hypothesis verification, thereby achieving the expansion from computational intelligence to memory intelligence.

The underlying logic of this trend is:The ultimate value of AI lies in simulating and augmenting human intelligence, and its evaluation system must take human cognition as the frame of reference. Cognitive formalization lays the foundation for constructing a new generation of artificial intelligence architecture that is interactive, learning-capable and self-evolving. Only evaluation grounded in cognitive science can truly reveal the capability boundaries and potential risks of AI systems, and achieve the fundamental leap from "measuring performance" to "measuring intelligence".

From the perspective of core theories in cognitive science, the essential necessity of this transition stems from the dual-process theory of human cognition (System 1 "fast intuition" and System 2 "slow rationality" proposed by Kahneman): The current mainstream AI evaluation systems mostly focus on testing the model's System 1 capabilities—intuitive and automated pattern matching based on training data—while lacking effective assessment of System 2 capabilities: high-level cognitive abilities such as slow thinking, logical reasoning, causal judgment, and reflection and revision. This is also the core reason for the phenomenon of "high scores but low ability" in models: models can accomplish tasks on static test datasets through pattern matching, but fail to complete tasks requiring logical reasoning and causal judgment in real complex scenarios. The core of the "Cognitive Science" evaluation paradigm is to reconstruct the theoretical foundation of AI evaluation with human cognitive mechanisms as the core frame of reference, and realize the fundamental shift from "result-oriented performance testing" to "process + result dual-oriented assessment of the essence of intelligence".

(II) Global Practice: Integrated Exploration of Cognitive Science and AI Evaluation

Leading global institutions have successively laid out strategies in this direction, forming diversified exploration paths. Through innovative theoretical frameworks, they are building quantifiable evaluation systems based on cognitive science:

In October 2025, Turing Award laureate Yoshua Bengio, together with scholars from 29 leading research institutions worldwide including Stanford University, MIT, and the University of California, Berkeley, published

a paper titled “A Definition of AGI” .This study established the first quantifiable evaluation framework for AGI, defining AGI as “AI that matches or exceeds well-educated human adults in cognitive diversity and proficiency” . Drawing on the authoritative Cattell-Horn-Carroll (CHC) theory of cognitive abilities in psychology, it decomposes general intelligence systems into ten core cognitive domains, including common sense and knowledge, literacy, mathematical ability, on-the-fly reasoning, and working memory, enabling modular, quantifiable assessment of AI cognitive capabilities. Tests based on this framework were conducted on GPT-4 (2023) and GPT-5 (2025). The results show that GPT-4 achieved an AGI score of 27%, while GPT-5 scored 57%. This marks the first time a standardized cognitive framework has been used to quantify the core gap between current large general purpose models and human cognitive abilities.

In 2025, a team led by scientist Anna A. Ivanova from the Georgia Institute of Technology published a paper entitled “How to Evaluate the Cognitive Abilities of LLMs” in “Nature Human Behaviour” , a sub-journal of Nature.The paper proposes an evaluation methodology covering 14 cognitive abilities: language comprehension, working memory, attention control, causal reasoning, analogical reasoning, theory of mind, metacognition, commonsense reasoning, moral reasoning, creativity, problem-solving, decision-making, spatial cognition, and numerical cognition.It emphasizes designing more robust evaluation protocols by simulating human cognitive processes to accurately interpret the core findings of research on AI cognitive ability assessment.

Dr. Nghia Duong-Trung from the German Research Center for Artificial Intelligence (DFKI), one of Germany’ s top AI research institutions, presented a paper entitled “BloomLLM: Large Language Models Based Question Generation Combining Supervised Fine-Tuning and Bloom” at the ECTEL 2024 conference. The paper introduces a method for fine-tuning large language models based on Bloom’ s Taxonomy of Cognitive Objectives, which includes the levels: remembering, understanding, applying, analyzing, synthesizing, evaluating, and creating. This represents an exploration of integrating cognitive science methodologies with large language models.

In China, in February 2025, the Research Institute of China Telecom, together with the Shanghai Institute of Innovative Algorithms, published a paper entitled “Attention Heads of Large Language Models” in Patterns - Cell. Drawing on methods from cognitive neuroscience, it proposed an innovative human brain-like cognitive framework consisting of four stages: knowledge recall, context recognition, latent reasoning, and expression preparation, aligning the reasoning process of LLMs with human reasoning mechanisms[figure 3]. Dr. Liu Quanying from Southern University of Science and Technology published a paper entitled “Promoting interactions between cognitive science and large language models” in The Innovation - Cell Press partner journal. He proposed using methodologies from cognitive science to evaluate the intelligence, mental states, and ethical levels of LLMs, providing theory and methodology for the multi-dimensional intelligent evaluation of large language models.

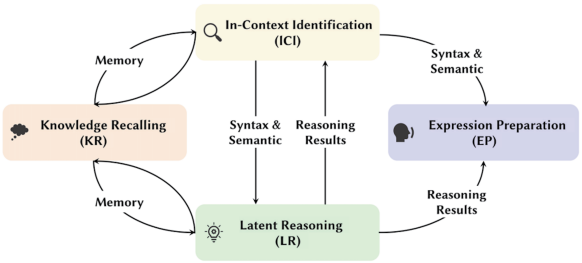

Figure 3: Cognitive Science Framework

(III) Core Value: Solving the Fundamental Problem of General Intelligence Evaluation

In response to this trend, the Research Institute of China Telecom has proposed the "Cognitive Science" evaluation paradigm. At its core, this paradigm deeply integrates cognitive science theory with AI evaluation. By aligning with human cognitive mechanisms, it drives the shift of evaluation from result-oriented to the dual orientation of "process + result", assesses the cognitive capabilities of large models across the entire input-processing-output chain, and achieves the precise measurement of the essential nature of AI intelligence.

The value of the "Cognitive Science" evaluation paradigm lies in three aspects: Firstly, it solves the "black box" problem by simulating human cognitive processes, making AI decision-making logic interpretable and traceable. Secondly, it enhances the foresight of evaluation, predicting AI performance in complex scenarios based on cognitive laws rather than relying on static datasets. Thirdly, it unifies the evaluation yardstick for general intelligence, providing a comparable and collaborative measurement system for global AGI development.

From the perspective of industrial deployment and security governance, the core value of the "Cognitive Science" evaluation paradigm is further extended: At the industrial level, AI models optimized based on cognitive alignment can better adapt to complex industrial business scenarios, especially high-value scenarios requiring logical reasoning, causal judgment, and creative problem-solving, greatly enhancing the business value and user satisfaction of AI applications. At the security governance level, cognitive alignment serves as the core foundation for AI safety alignment. Only by achieving alignment between AI and human cognitive mechanisms can we fundamentally align AI with human values and ethical norms, and prevent catastrophic risks of AI.

III、Trend 2: Scenario Deepening – From Generic Benchmarks to Precise Penetration in Vertical Domains

(I) Trend Connotation: Industrial Deployment Forces the Scenario-based Transformation of Evaluation

The adoption rate of enterprise-level AI applications is rising year by year, yet it remains difficult to sustainably deliver quantifiable business value to enterprises. The core bottleneck lies in the lack of a full-process evaluation system adapted to industry-specific scenarios. One of the key pain points in current AI evaluation is the disconnect between general benchmarks and real industrial needs — high scores achieved by models on general benchmarks such as MMLU and GLUE can hardly be translated into practical value in specific industries. As AI technology moves from the laboratory to industry, evaluation systems must penetrate deeply into vertical scenarios. Refined and customized evaluation frameworks should be built according to the business logic, data characteristics, and security requirements of different industries. The core of this trend is "boosting deployment through evaluation", using scenario-based evaluation to bridge the last distance between AI technology and industrial value.

The core essence of scenario-based evaluation is the shift from model-centric general performance testing to business-centric full-dimensional value assessment, which follows four core principles:

a.Business-oriented principle: All evaluation indicators are designed around core industry business objectives, rather than generic technical metrics.

b.Real-environment principle: Evaluation data and test scenarios fully replicate real industry business environments, including complex factors such as noisy data, edge cases, and unexpected situations.

c.Full-lifecycle principle: Evaluation runs through the entire process of AI application, including requirement design,model training, online deployment, and operation & maintenance, rather than being a one-time acceptance check before launch.

d.Risk-forwarding principle: Identify compliance risks, security risks, and business risks in scenarios in advance during evaluation, achieving early detection and early handling of risks.

(II) Global Practice: Diversified Exploration of Industry-Tailored Evaluation

Mature practices have been formed worldwide across multiple high-value sectors, confirming the necessity of scenario-based evaluation:

For high-risk fields:

a.In medical scenarios: Evaluation focuses on diagnostic accuracy, cross-device generalization, and clinical interpretability. For example, the U.S. FDA's approval requirements for medical AI include specificity and sensitivity metrics supported by multi-center clinical trial data.

b.In financial scenarios: Evaluation of anti-fraud models emphasizes not only prediction accuracy, but also fairness (avoiding discrimination against specific groups), adversarial robustness (defending against evolving fraud techniques), and regulatory compliance (meeting anti-money laundering and data privacy requirements).

For vertical industry fields:

a.In customer service scenarios, China Telecom has built an evaluation system for the customer service industry. It designs targeted evaluation metrics around business scenarios such as script recommendation, report generation, and call summary, enabling deep integration of evaluation into business processes and helping improve the productivity of AI applications in industry scenarios.

b.In government affairs scenarios, targeting

demands such as social security and household registration consultation, evaluation incorporates full-process metrics including “user intent understanding, policy knowledge reasoning, and compliant expression”, ensuring that AI applications are both efficient and secure.

In addition, the “Huiju Zhiping” (Smart Gathering & Intelligent Evaluation) Large Model Evaluation Working Group, jointly led by China Telecom, China Mobile, and the China Electronics Standardization Institute, has achieved a scale effect in the field of industry-specific large model evaluation. It has established systematic evaluation standards for industries including electric power, logistics, petrochemical, and transportation. Relevant enterprises are cooperating to carry out evaluation implementation. For example, in 2025, evaluations were conducted for 5 enterprises in the electric power industry, deeply integrating ecological forces.

For future industrial layout, scenario-based evaluation will take an early lead in standard-setting for emerging fields such as embodied intelligence, autonomous driving, and low-altitude economy. For instance, evaluation for autonomous driving has shifted from closed-field testing to real-road "field" evaluation, covering complex scenarios including severe weather and emergencies.China Telecom has established scenario-based evaluation indicators for humanoid robots, intelligent drones and other platforms, focusing on environmental perception, path planning, and emergency decision-making, laying a foundation for the development of future industries.

(III) Core Value: Accelerating the Large-Scale Deployment of the AI Industry

The core value of scenario-based evaluation lies in “precision matching”:It provides enterprises with technical selection basis that fits their own needs, reducing trial-and-error costs; It offers developers clear optimization directions, preventing a disconnect between technical R&D and market demand; It supplies regulators with domain-specific risk assessment tools, enabling “precision governance”. Globally, scenario-based evaluation has become an important indicator of AI industry maturity. Its popularity directly determines the speed and depth of AI technology penetration into all industries.

From the underlying logic of both supply and demand in the industry, the core value of scenario based evaluation lies in breaking the structural mismatch between generic evaluation systems and real industrial needs, driving a fundamental shift of AI evaluation from “technology oriented” to “value oriented”. For technology demanders:By anchoring real business objectives to build evaluation benchmarks, scenario based evaluation eliminates the information gap between generic technical metrics and actual business value.It enables market entities to break free from selection misconceptions such as “parameter centricism” and “ranking centricism”, and establish a selection and decision making system centered on business value. This fundamentally reduces the trial and error and decision making costs of AI deployment.It also provides equal technical selection references for market entities of different scales and with varying digital foundations, removing cognitive and decision making barriers for the inclusive application of AI technologies.

For technology suppliers, scenario-based evaluation converts vague and scattered industrial demands

into quantifiable, implementable technical optimization objectives, completely resolving the industry pain point of disconnection between R&D and market demand. This evaluation system can redirect R&D resources away from a pure competition in parameter scale and score-chasing on generic benchmarks, toward technological innovation that genuinely solves real industrial problems, achieving a two-way improvement in R&D efficiency and industrial value. Meanwhile, scenario-based evaluation establishes a clear feedback loop for AI technology iteration that aligns with industrial needs, ensuring that technical optimization always revolves around real industrial demands and driving the steady, continuous evolution of AI technologies from laboratory innovation to industrial application.

On the governance side, scenario-based evaluation serves as the core vehicle for achieving tiered and categorized precise governance of AI, driving the transformation of AI governance from a one-size-fits-all universal regulatory framework to refined governance adapted to the characteristics of different sectors. By establishing differentiated evaluation systems tailored to the business logic, risk levels, and compliance requirements of various industries, scenario-based evaluation translates macro-level ethical principles, safety norms, and compliance mandates into concrete, operable, verifiable, and traceable assessment indicators. This not only accurately identifies potential risks of AI systems in different scenarios to firmly uphold the bottom line of safe development, but also avoids the constraints of excessive regulation on technological innovation. In this way, it achieves a dynamic balance between the development and security of the AI industry, providing solid tool support for the implementation of the global AI governance system.

From an overall industrial development perspective, the maturity of scenario based evaluation is a core indicator that the AI industry has shifted from the technological exploration stage to the stage of large scale deployment. When the AI evaluation system completes its deep penetration from generic benchmarks to vertical scenarios, it signifies that the integration of AI technology and the real economy has entered a new phase of systematization and standardization. Scenario based evaluation can establish unified industrial consensus and value benchmarks for the deployment of AI technology across all industries, break down technical and information barriers between different sectors and stakeholders, and accelerate the penetration of AI technology from standalone applications to the entire industrial chain and business processes. The depth of its adoption and degree of standardization in various industries not only determine the speed and scope of AI technology integration into the real economy, but also fundamentally define the overall development quality and sustainable growth potential of the AI industry.

IV、Trend 3: Ecological Collaboration – Dual Drivers of Platform-based Support and Governance Upgrade

(I) Trend Connotation: Systematic Evolution from a Single Tool to a Collaborative Ecosystem

AI evaluation is growing increasingly complex, and single institutions or individual tools can no longer meet the demands of the entire industrial chain. The future evaluation system will feature the dual characteristics of platform-based support + governance upgrade:

a.Platformization addresses issues of usability and scalability, lowering evaluation barriers through one-stop, end-to-end tools that cover all users, from ordinary users to developers and regulators.

b.Governance upgrade ensures credibility and uniformity, establishing a transparent, independent, and accountable governance framework to guarantee the objectivity and authority of evaluation results.

These dual characters form an open and collaborative global AI evaluation ecosystem together. The underlying logic of this trend is:AI evaluation has become a factor-based tool for global governance and must itself have a sound governance system. Meanwhile, the industry' s demand for low-cost and large-scaled evaluation is driving the platform integration of tools.

From the perspective of industrial chain division of labor, the AI evaluation ecosystem has formed six clear core links:

a.Evaluation theory and standard research institutions: responsible for building the theoretical framework and standard system of AI evaluation.

b.Evaluation dataset providers: responsible for developing standardized and scenario-based evaluation datasets.

c.Evaluation tool and platform developers: responsible for researching and developing evaluation tools, building evaluation platforms, and realizing the engineering and large scale implementation of evaluation.

d.Third party independent evaluation institutions: responsible for providing objective and impartial third party evaluation services and issuing evaluation reports.

e.Industrial application parties: including governments, enterprises and other users of AI systems, who are the core demanders of evaluation services.

f.Regulators: responsible for formulating regulatory rules for evaluation and standardizing the development of the evaluation industry.

Only through the collaboration of the six links can a complete, healthy and sustainable AI evaluation ecosystem be built.Platformization is the core carrier of ecological collaboration, and governance is the core guarantee of ecological collaboration. The two are complementary and indispensable.

(II) Global Practice: Parallel Progress of Platform Development and Governance Framework

In terms of platform-based practices: Internationally, OpenAI has launched the Evals open-source evaluation framework, which provides researchers and developers with standardized evaluation tasks and architectures to compare the performance of different large language models (LLMs) across various dimensions. Companies including Google (Vertex AI) and Amazon (Bedrock) have introduced MaaS platforms with built-in evaluation tools, realizing the integration of model deployment and evaluation. In China, in addition to platforms such as Sinan that focus on building influence through rankings, independent platforms dedicated to full-stack AI evaluation capabilities have gradually emerged. For example, China Telecom' s "Tiangang" AI Evaluation Platform [figure 4]features "one-stop, end-to-end, and visual" core functions. It tracks and adapts to mainstream large models worldwide and builds proprietary evaluation datasets, covering more than 200 general

models[Figure 5], 40 industry models, and 100 segmented scenarios. The platform supports all user groups, from ordinary users (visual operations) to developers (private deployment or API interfaces) and regulators (special modules for security and compliance), lowering the threshold for using evaluation services.

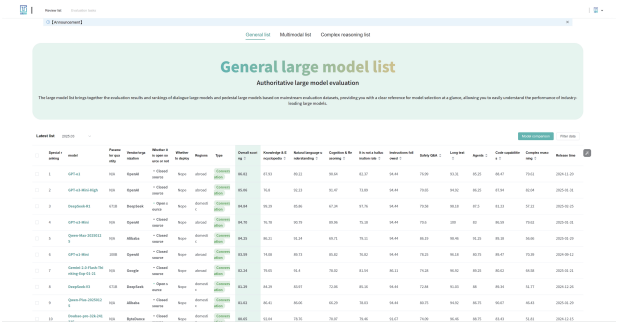

Figure 4: General Large Model Ranking List on Tiangang AI Evaluation Platform

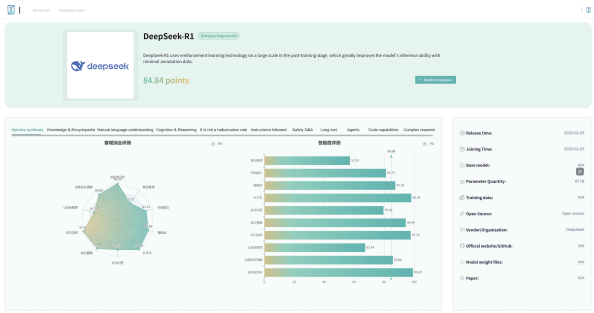

Figure 5: Evaluation Result Details of DeepSeek-R1 on Tiangang AI Evaluation Platform

In terms of governance exploration, the world is focusing on the core issue of "who evaluates the evaluators" to build an evaluation governance framework. The EU AI Act explicitly requires that institutions conducting high-risk AI evaluations must maintain independence and transparency. Academic research has extended the "evaluation card" mechanism to the evaluation field, advocating that evaluators and model developers disclose information such as design principles, data composition, and potential biases. The World Economic Forum (WEF) is promoting the establishment of a global AI evaluation governance framework to facilitate cross-national and cross-industry standard mutual recognition. Against the backdrop of China's AI development context, it is also necessary to cultivate localized governance solutions. For instance, in its evaluation practices, China Telecom has incorporated ideological alignment and value alignment as core indicators, forming a compliant evaluation module that meets China's governance requirements, and providing a practical local example for governance-oriented evaluation.

(III) Core Value: Building a Trustworthy and Inclusive Global AI Ecosystem

The synergy between platformization and governance will deliver two major values: Firstly, inclusiveness, enabling more enterprises and institutions to access professional evaluation services through low-threshold tools, thus narrowing the "digital divide". Secondly, credibility, enhancing the global recognition of evaluation results through a transparent and independent governance framework, and breaking the fragmented situation of "benchmark manipulation" and "uncoordinated narratives". Ultimately, this will build an open, collaborative, trustworthy and reliable global AI evaluation ecosystem, providing systematic support for the healthy development of AI technologies.

From the perspective of global AI governance, the core value of ecological collaboration lies in reducing compliance barriers to the cross-border flow of AI technologies through globally unified evaluation standards and governance frameworks, and promoting collaborative innovation and the healthy development of the global AI industry.As a global technology, AI requires global collaboration in its governance, and AI evaluation serves as the key instrument for implementing global AI governance.Only by achieving mutual recognition of global evaluation standards can we coordinate global AI governance rules and avoid the formation of "digital

barriers" and "governance fragmentation" . From the perspective of AI security, an ecologically collaborative evaluation system can pool global wisdom and resources to identify potential risks of cutting-edge AI systems at an earlier stage, build a globally unified line of defense for AI security, prevent the abuse and catastrophic risks of AI technologies, and ensure the secure and controllable development of AI. From the perspective of inclusive development, platform-based evaluation tools can significantly lower the technical threshold and cost of AI evaluation, enabling small and medium-sized enterprises and developing countries to access professional evaluation services on an equal footing.This narrows the AI capability gap between developed and developing countries, and between large enterprises and SMEs, promotes the inclusive application of AI technologies, and fulfills the core goal of "AI for Good".

V、Challenges and Recommendations for the Development of Global AI Evaluation

(I) Core Challenges

1.Evaluation Technology Level

Cognitive Evaluation Technology Bottleneck: Quantitative evaluation methods for high-level cognitive abilities such as common-sense reasoning, causal judgment, and creativity remain immature. Existing technologies struggle to comprehensively measure the essential intelligence of AI.

Lagging Capabilities in Cutting-edge Technology Evaluation: Cutting-edge technologies such as AGI, embodied intelligence, multi-modal AI agents, and on-device AI are accelerating iteration. However, the development of corresponding evaluation theories, methods, and tools is relatively lagging behind, resulting in a situation where technologies advance first while evaluation follows behind, which fails to provide effective support for the secure and controllable development of cutting-edge technologies.

2.Scenario Implementation Level

Scenario-based Data Barriers: Real data in vertical industries often involves privacy or trade secrets, making it difficult to share, which restricts the development of scenario-based evaluation.

Insufficient Integration of Evaluation into the Whole R&D Process: At present, AI evaluation in most enterprises still remains at the "one-time acceptance" stage before deployment. Evaluation has not been integrated into the full-lifecycle of AI R&D, deployment, and operation. As a result, risks during the R&D process cannot be identified in advance, and model optimization lacks precise guidance, leading to the problem of disconnection between evaluation and R&D.

3.Ecological Collaboration Level

Standard Fragmentation: The lack of unified evaluation standards on a global scale, along with large discrepancies in benchmarks across institutions and countries, makes it difficult to mutually recognize evaluation results, increasing compliance costs for enterprises and complicating the cross-border flow of technologies.

Lack of Fairness and Inclusiveness in Evaluation: Currently, most of the world’ s mainstream evaluation benchmarks and datasets are based on the English language and Western cultural backgrounds, resulting in severe inadequacy in adapting to non-English languages and local scenarios in developing countries. This leads to systemic biases in evaluation results and exacerbates the digital divide in global AI development.

Governance System Imperfection: The lack of globally unified norms for the independence, transparency, and accountability of evaluation institutions may lead to the abuse of evaluation results and undermine their credibility.

(II) Recommendations for the Development of AI Evaluation

1.To Policymakers: Promote Standard Mutual Recognition and Governance Collaboration

1.Strengthen international cooperation, lead or participate in the formulation of global AI evaluation standards, promote the mutual recognition of core indicators based on cognitive science, and break the pattern of fragmented standards.

2.Increase financial support for the R&D of core AI evaluation technologies, such as establishing special research funds to support the development of core evaluation technologies for emerging fields including cognitive alignment evaluation, scenario-based evaluation, AI security evaluation, and embodied intelligence, so as to

break through technical bottlenecks.

3.On the premise of ensuring data security and privacy, promote the open sharing of public data and high-quality industry datasets to provide data support for scenario-based evaluation.

4.Cultivate third-party independent evaluation institutions, establish a sound qualification certification system, supervision norms and exit mechanism for evaluation institutions, regulate the development of the evaluation industry, and improve its overall professional level and credibility.

5.Establish a supervision framework for evaluation institutions, clarify requirements for independence and transparency, and ensure that the evaluation process is traceable and supervisable.

2.To Industry: Embrace Scenario-based and Platform-based Development, and Practice Responsible Evaluation

1.Actively participate in the co-construction of scenario-based evaluation standards, transform business experience into universal industry indicators, and promote the deep integration of evaluation with industrial demands.

2.Strengthen industry-university-research-application collaboration with research institutions and universities, jointly conduct theoretical research, technological development and standard-setting for AI evaluation, and drive the innovation and implementation of evaluation technologies.

3.Carry out full-lifecycle evaluation based on platform-based tools, integrate evaluation into the entire process of AI product design, development and deployment, and avoid risks in advance.

4.Establish an internal AI evaluation governance system, clarify responsible departments, process specifications and quality standards for evaluation, take AI evaluation as a mandatory prerequisite for AI product launch, and fulfill the primary responsibility of enterprises.

5.Adhere to the principle of transparency, disclose the evaluation methods and results of AI products, refrain from speculative behaviors such as “benchmark gaming”, and jointly safeguard the credibility of the evaluation ecosystem.

3.To Research Institutions: Deepen Interdisciplinary Research and Break Through Technical Bottlenecks

1.Focus on evaluation research for cutting-edge technologies. For AGI, embodied intelligence, AI agents and other advanced technologies, advance the layout of research on evaluation theories and methods to achieve synchronous development of evaluation and technological innovation.

2.Strengthen the interdisciplinary integration of AI evaluation with cognitive science, psychology, ethics and law, explore quantitative evaluation methods for high-level cognitive abilities, and improve the "cognition +" evaluation paradigm.

3.Develop dynamic and interactive evaluation technologies to simulate the complexity and uncertainty of the real world, and bridge the gap between "laboratory performance" and "practical application effectiveness".

4.Pay attention to the social impact of evaluation, research indicators related to fairness and inclusiveness, and promote the development of AI technologies toward "AI for Good".

5.Strengthen international academic exchanges and cooperation, participate deeply in global research on AI evaluation theories and technologies, and promote collaborative innovation of global evaluation theories and methods.

4.For Third-Party Evaluation Institutions: Uphold Independence and Impartiality, Enhance Professional Competence

1.Adhere to the core principles of independence, objectivity and impartiality, establish a sound internal governance system and conflict-of-interest isolation mechanism, eliminate interest ties with evaluated entities, and ensure the credibility of evaluation results.

2.Continuously strengthen technical capacity building, keep pace with the iterative evolution of AI technologies, steadily enhance evaluation capabilities for cutting-edge AI technologies, and provide professional, authoritative and comprehensive evaluation services for industries and regulators.

3.Strictly comply with AI regulatory laws and industry standards worldwide, standardize evaluation procedures, fully record information throughout the entire evaluation process, and ensure that the evaluation process is traceable, auditable and reproducible.

5.To International Organizations: Promote Global Collaboration and Advance Inclusive Development

1.Leverage the coordinating role of international organizations, build a global collaboration platform for AI evaluation standards, promote the unification and mutual recognition of global AI evaluation standards, and address the core issue of standard fragmentation.

2.Establish a global sharing platform for AI evaluation resources, promote worldwide sharing of evaluation datasets, tools, methodologies and computing resources, and narrow the gap in AI evaluation capabilities between developed and developing countries.

3.Promote the coordination of global AI evaluation governance, formulate unified global principles and guidelines for AI evaluation governance, regulate the development of the global AI evaluation industry, and enhance the global credibility of AI evaluation.

VI、Conclusion

The future development of AI evaluation will center on the three core trends of cognitive alignment, scenario deepening, and ecological collaboration. This is not only an inevitable outcome of technological evolution, but also an objective requirement for the sound development and effective governance of the global AI industry. The "Cognitive Science" paradigm strengthens the theoretical foundation of evaluation, achieving a leap from "measuring performance" to "measuring intelligence"; scenario-based evaluation connects technology and industry, accelerating the large-scale application of AI; platform-based and governance collaboration builds a credible and inclusive ecological system, providing safeguards for the development of global AI.

In the long run, the development of AI evaluation will profoundly shape the direction, industrial landscape and governance rules of global AI technology. Against the backdrop of the accelerated evolution of artificial general intelligence, AI evaluation is no longer a mere "supporting tool" for AI technology. Instead, it has become core infrastructure that determines the secure and controllable development and large-scale industrial application of AI, as well as a commanding height in the global competition for AI technology and governance discourse power.

Faced with global competition and governance challenges, countries and institutions need to uphold an open and collaborative attitude to jointly promote the mutual recognition of evaluation standards, technological innovation and improved governance. Practices of global organizations have provided an initial solution of "cognitive alignment + scenario implementation + platform support" for global AI evaluation. In the future, with the in-depth development of the three major trends, through continuously deepening theoretical innovation of the "Cognitive Science" evaluation paradigm, promoting the industry-wide popularization of scenario-based evaluation, and building an open and collaborative evaluation ecosystem, AI evaluation will truly become the compass for global AI technological innovation, the yardstick for industrial implementation, and the cornerstone of governance rules, providing solid support for building an intelligent era of human-machine coexistence.