CAICT 中国信通院

Deepen Intelligent Evolution and Build a New AI Landscape: AIIA Advances with the Industry

Feng Cao
Institute of Artificial Intelligence, CAICT

March 2026

An inflection point for AI: From technological innovation to in-depth industry application

CAICT 中国信通院

~2024

Expansion

Technology: Technical paradigms established

Scaling laws and paradigms established

Application: Implementation of pilot projects and exploration, and adaption of technologies to industry scenarios

AI-based quality inspection in manufacturing, intelligent Q&A in public services, risk control in finance, AI-assisted diagnosis in healthcare, etc.

2025

Inflection point

Technology: Evolution of core technologies

Foundation models → Thinking models

Chat with users → Execute tasks

Generally closed-source → The rise of open source

Application: Large-scale and in-depth implementation to improve quality and efficiency in all industries

Knowledge base, BI, code generation, text generation, etc.

2026~

Breakthroughs

Technology: Continued expansion and architecture innovation

Focus on cutting-edge technologies (e.g., agentic AI, agent engineering, Skills, and AI-native).

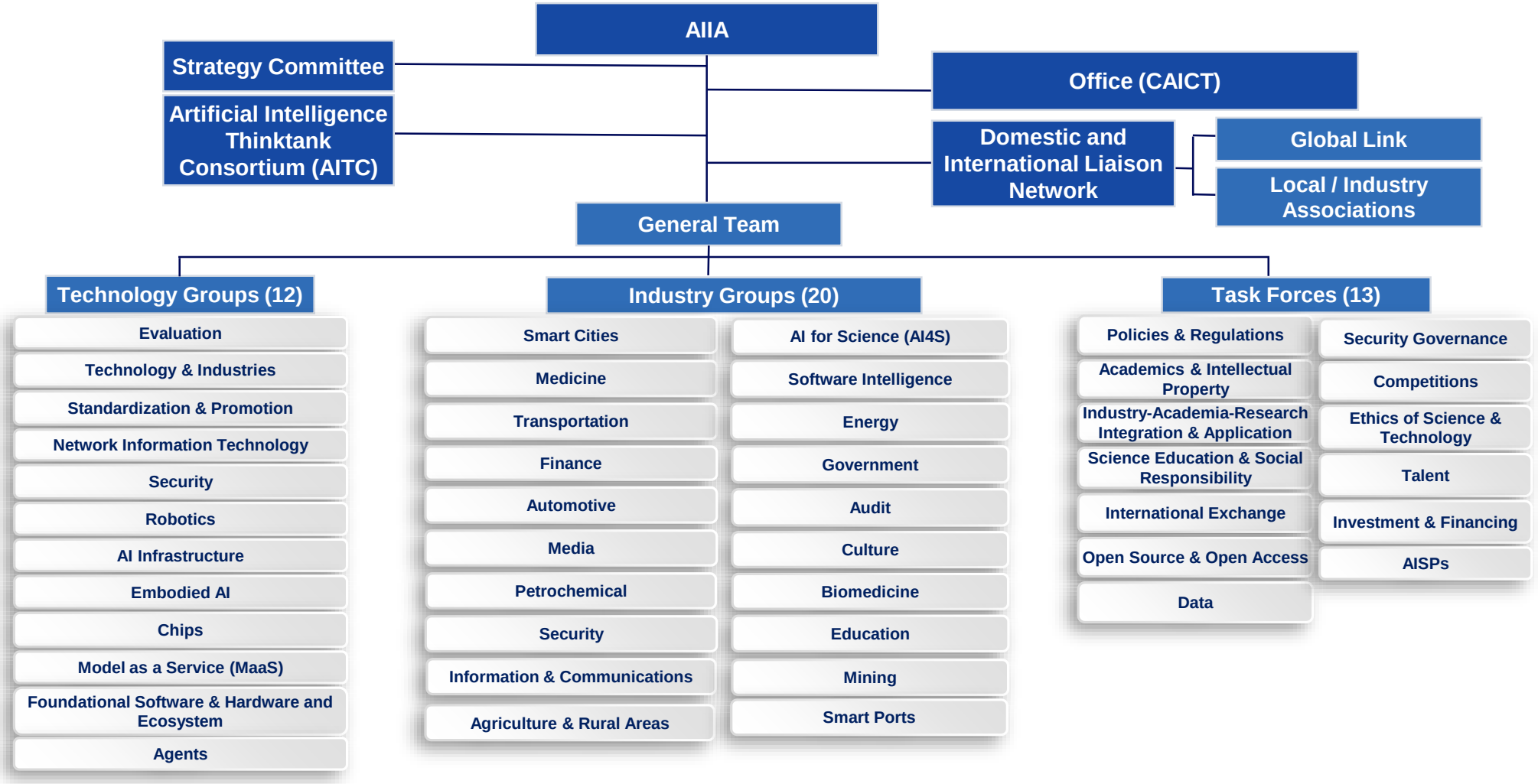
Application: Extensive adoption of agents and iterative ecosystem upgrade

Autonomous agents, smart workforce, scientific research, innovation, etc.

About the AIIA: *Empower, collaborate, and lead*

CAICT 中国信通院

- **China's Artificial Intelligence Industry Alliance (AIIA) is an industry collaboration mechanism jointly initiated by CAICT and many other organizations in 2017. AIIA currently has 1,439 member organizations and 45 working groups.**



Positioning

Drive the wide adoption of AI to empower digital economy growth.

Vision

Become a public service platform for AI development that has global influence.

Mission

Technological innovation, application promotion, industry ecosystem cultivation, and international collaboration

Key tasks of the AIIA

➢ **With a focus on the key directions of the AI industry, the AIIA aims to drive the high-quality development of this industry by arranging for its members to work on tasks such as making technical breakthroughs, empowering application, aligning supply and demand, developing standards, and conducting international exchanges and collaboration.**

Technological innovation & evaluation

- Establish multi-dimensional evaluation systems, build the **FacTesting** benchmark test system, and develop multiple types of evaluation systems, such as those targeting AI dataset quality, embodied AI, and agents.
- Run the **AITC** to analyze trends, discuss technologies, and support policies.

Industry application & ecosystem cultivation

- Build **collaboration platforms** to align supply and demand and host high-level competitions.
- Build a **high-value scenario library** to drive the wide adoption of AI in industries.
- Create **local platforms**, build public service platforms for large models, and assist in building data evaluation centers in multiple locales.

Security governance & open collaboration

- Drive the signing and verification of the **AI Security Commitment**, and conduct research on agent security standards.
- Organize participation in international activities (e.g., **WAIC and AI for Good Global Summit**) to promote China's success stories in AI worldwide.
- Drive **industry-wide collaboration for breakthroughs**, and organize chip-model linkage and the cultivation of the open-source intelligent computing ecosystem.

Creating the "data-model-application" flywheel to deepen implementation in scenarios

CAICT 中国信通院

Value of AI implementation

Enterprises actively explore the value of AI

Higher work efficiency

Gaining insights into knowledge and data

Production tool empowerment

Management model innovation

Internal empowerment

Implementation of large models

External services

Business guidance and efficiency improvement

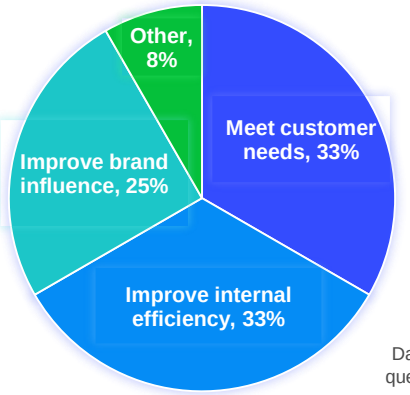
Product and service upgrade

Marketing tool empowerment

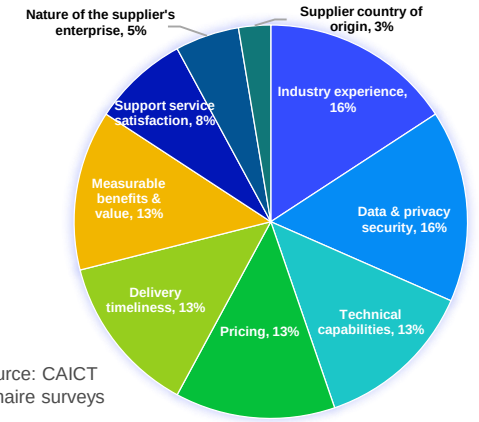
Market strategy decision making

Evaluation of the results of implementation

Drivers behind enterprises' use of large models



Factors affecting enterprises' purchase of large models

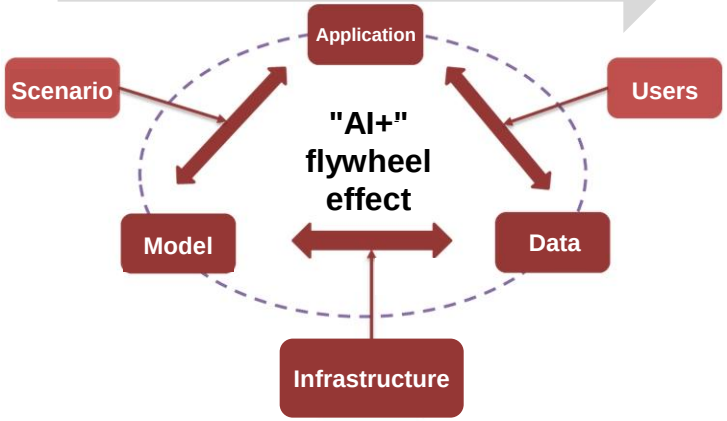


Data source: CAICT questionnaire surveys

Drivers behind the use of large models: **Meeting application requirements** and **improving internal efficiency**
Factors affecting purchase: Industry experience and **data & privacy security**

Results

"Data-model-application" flywheel



2. Model → Application: Drive the in-depth integration of models and scenarios.

3. Application → Data: Deeply explore high-quality data resources through application.

Support elements

Resource-to-technology conversion

Application-to-resource conversion

1. Data → Model: Consolidate the resource foundation and convert the advantages in data resources into the advantages in intelligent capabilities.

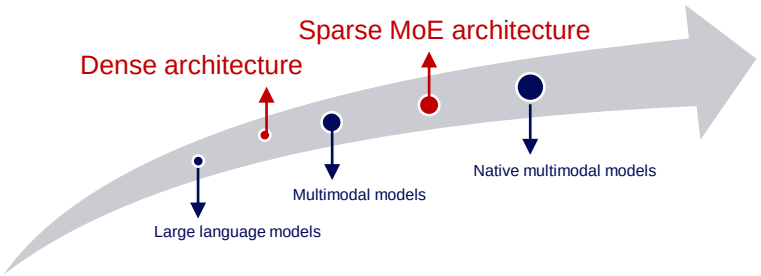
Enterprise-grade AI infrastructure: Expanding the focus from "usable" to results, performance, and costs

CAICT 中国信通院

➢ **The large-scale implementation of large models drives the in-depth optimization of the inference architecture and enables the achievement of goals for results, performance, and costs.**

Adaptation challenges from multiple dimensions

1. Models: Multimodal and MoE models become the mainstream



2. Application: Requirements for long context and memory

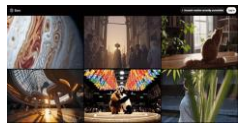
Continuous development of **RAG and agent services** and explosive growth of new requirements
→ **Requirements for long context**
→ **Requirements for memory**

Long text + Memory

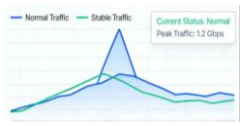
Parameterized memory	Context memory
External databases	Memory OS

3. Scenarios: Low latency, high throughput, and traffic fluctuations

Low latency
Intelligent customer service and dialog



High concurrency
Batch dataset building



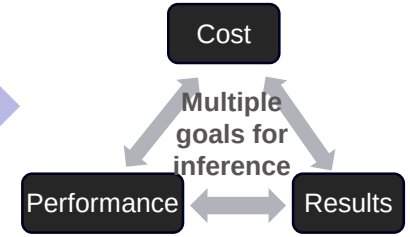
Traffic fluctuations
Peak hours in the morning & afternoon, at weekends, and during holidays

Achievement of multiple goals (results, performance, and costs) through reference optimization

Balance between computing power and cost

Current dilemma:
How can we maximize performance and reduce costs with limited computing power?

Simply stacking more chips for higher performance
→ **Increases costs**
→ **Makes it difficult to generate sustainable profits**



The major approach: Inference optimization

Single-instance optimization: Efficient request computing

Use **inference engines, tools, and components** as vehicles to connect hardware, software, and models.

- Model compression
- Computing optimization
- High-performance storage
- Multi-step reasoning
- Batch scheduling policy
- Parallelism policy

For scenarios with low request traffic and low computing power

Multi-instance optimization: Request and resource allocation

Consider all relevant factors such as **models, resources, and scenarios (e.g., long text, latency, and concurrency)**.

- Prefill-Decode disaggregation
- AF separation
- Instance fragment management
- MoE cluster
- KV cache pooling
- Adaptive scheduling policy

For scenarios with high service traffic and sufficient computing power

Related tasks for CAICT

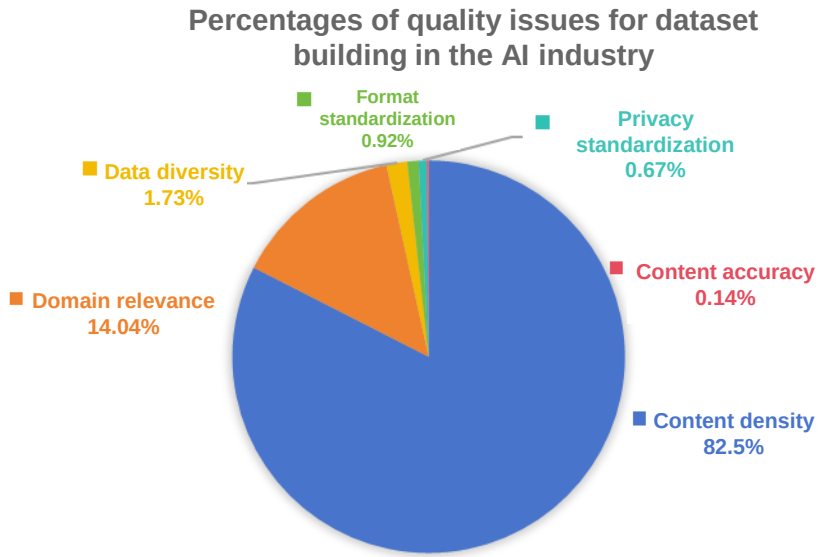
1. Rating criteria for inference optimization: Optimize the technical guide.
2. Inference test beds: Perform horizontal quantitative bottleneck localization.
3. Research reports: Provide system-level optimization solutions for reference and the optimization path guide.

Enterprise data governance: Focusing AI-ready data engineering on addressing bottlenecks for dataset implementation

CAICT 中国信通院

➢ **Dataset building: Methodology guidance and technological innovation are urgently needed, and there are four key types of data requirements in 2026.**

Quality issues and methodologies are the key to high-quality datasets



- Domain relevance
- Content density
- Data diversity
- Format standardization

From the perspective of enterprise practice, the focus of dataset building has shifted from general and basic data to industry scenarios. Dataset quality issues have become the bottleneck for vertical model implementation and scenario application. The following three major issues will hinder enterprise efforts to build high-quality datasets.

Goals and positioning not clearly defined

Scenario requirements are disconnected from dataset building goals, and data engineering goals are not closely linked to core business indicators.

Fragmented implementation paths

The entire link from data collection to model training is not systematically planned or designed, and there is no systematic dataset building and maintenance mechanism.

Weak technical foundation

There is no tool chain that adapts to industry characteristics. The level of automation is low, and there is heavy reliance on manual operations, hindering engineering implementation efficiency.

Future AGI development is highly dependent on high knowledge density and scenario data

1. Data requirements of world models

WorldScore understands and predicts the dynamic evolution of the physical world.

2. Data requirements of embodied AI

1% real-world data vs. 99% synthetic data

3. Data requirements of agents

Interaction data that is deeply coupled with the test environment

https://os-world.github.io/

4. Data requirements of industry models

Healthcare: Chain of thought data in the diagnosis process
Finance: Risk control and customer profile data
Transportation: Traffic flow and detection data
Industrial: Process knowledge and production process data
Energy: New energy station and power grid inspection data

Driving high-quality dataset building

1. AI-ready enterprise data engineering

Closed-loop management of the end-to-end process from **system building, development and maintenance, quality control, and resource operations to compliance and trustworthiness**

2. Driving breakthroughs in cutting-edge data technologies

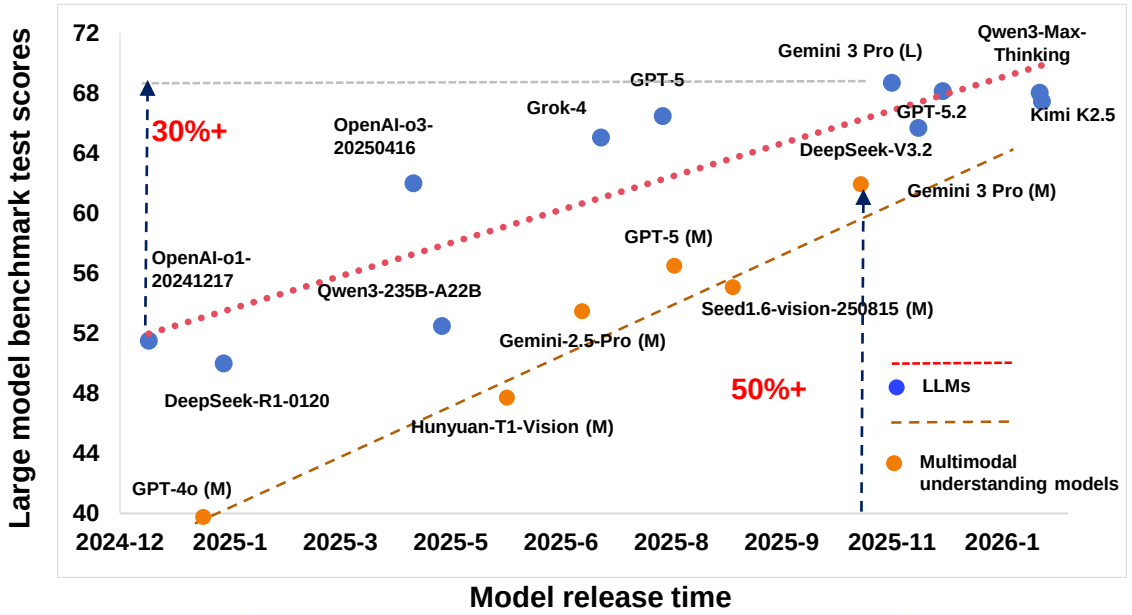
Encourage enterprises to work with universities and research institutes to make breakthroughs in technologies such as **industrial-scale production of synthetic data, data augmentation and expansion, full-link quality control, and heterogeneous data**.

Specialized models: Capabilities of large models have improved significantly, but key breakthroughs are needed in terms of industry application

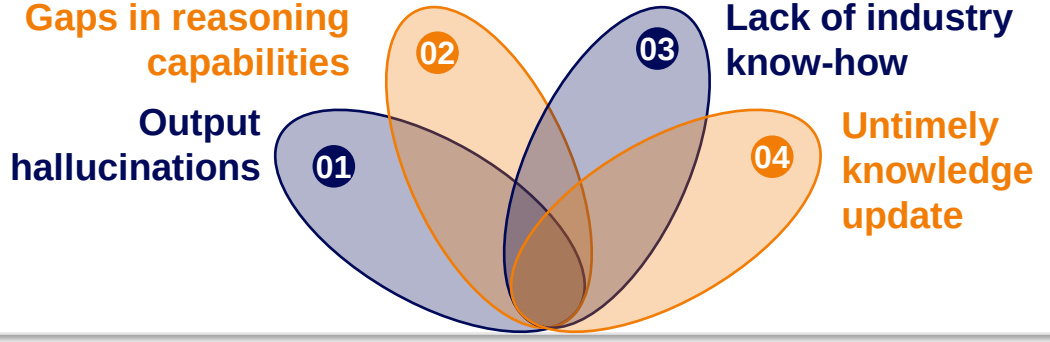
CAICT 中国信通院

➢ **In 2025, the capabilities of LLMs and multimodal models worldwide improved by 30%+, but there are still gaps in their industry application. Testing is a key means to drive the industry application of models.**

Significant progress in LLMs and multimodal models

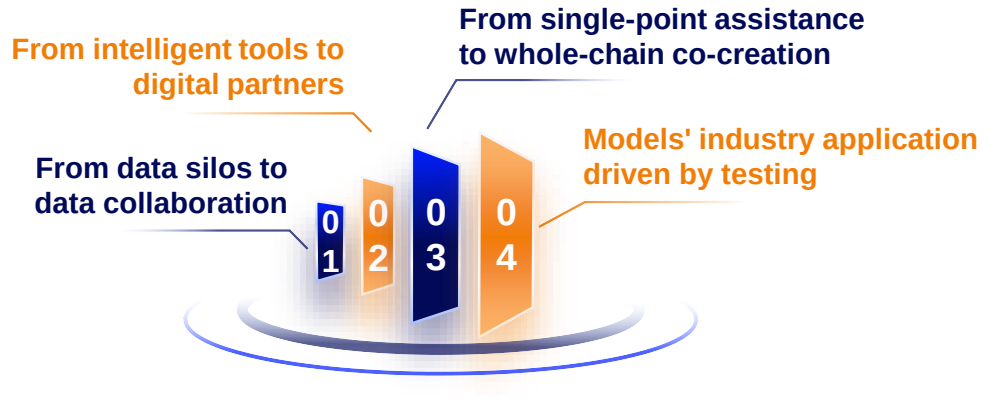


Models' industry application: Issues and challenges



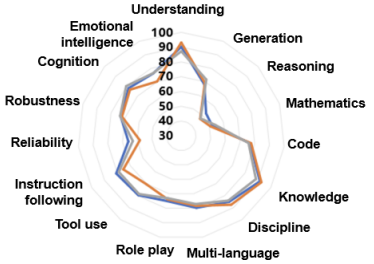
Large model benchmark test systems support model application

Future trends of models' industry application

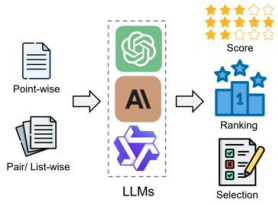


FacTesting large model benchmark test system

Test standards



Test data



Test technology

Test tools

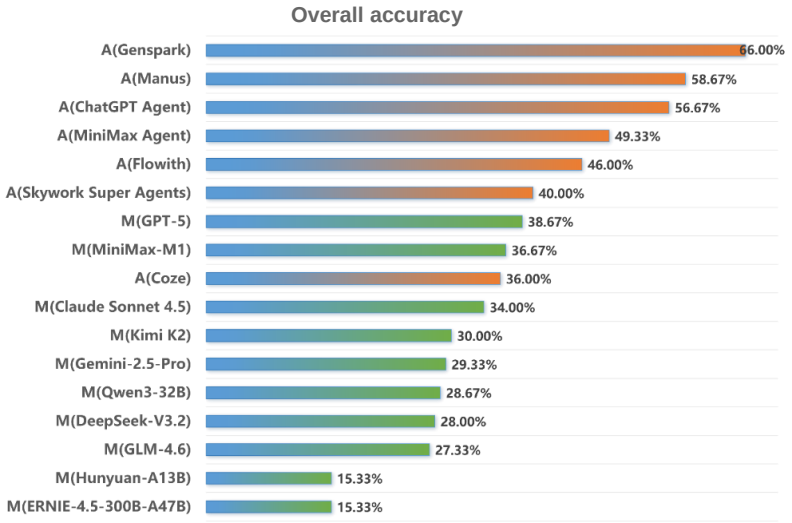
Agents: Greatly pushing the boundaries of AI capabilities as the main manifestation of models

CAICT 中国信通院

➢ **Agents are an engineering enhancement of large models, but their large-scale application still faces multiple challenges.**

Agents are becoming increasingly autonomous, but their application still faces challenges

- **Agents perform better than individual models, but their performance still falls short of expectations.**



Note: "A" stands for agents and "M" stands for models. Data source: FacTesting benchmark test by CAICT, February 2026

- **There are still challenges in the technical capabilities and engineering architecture of agents.**

Technical capabilities

Gaps in model capabilities
Gaps in handling complex long-horizon tasks
Gaps in multi-agent collaboration

Engineering architecture

Fast evolution of technologies such as Skills, MCP, and memory
Complex agent engineering architecture
Bottlenecks in memory and knowledge management
Gaps in tool invocation efficiency and accuracy

An urgent need to evolve from single-point agents to enterprise-grade agentic systems

- **Enterprise-grade agent architecture: Driving the evolution of single-point agents to complex agentic systems**

- **Integration with enterprises' existing IT and data infrastructure**
- **Integration with enterprises' business systems (e.g., ERP and CRM)**
- **Integration with enterprises' work processes (e.g., procurement, R&D, and O&M)**

- **Agent-related research work conducted by CAICT**

Standards development	Agent test system	Agent test beds
Agent standards system	Basic capability test	MCP development and test
International standards and industry standards	General task test	Single agent capability test
	Typical application scenario test	Multi-agent collaboration test

AI-native: Driving a paradigm shift in application and requiring ecosystem development

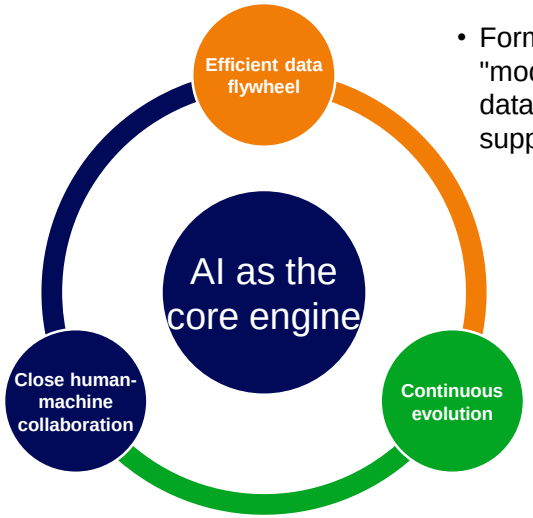
CAICT 中国信通院

➢ **AI-native is set to change the game, drive a paradigm shift in application, and evolve from concept to reality. It is now an urgent task to drive the development of standards, specifications, and test capabilities on AI-native, and support the development of AI-native technologies and the industry ecosystem.**

AI-native is reshaping products, services, and how enterprises are organized

- **AI-native: Use AI as the core engine to change the game and reshape technologies, products, business models, and operations management.**

Features of AI-native



- Form an efficient "model-scenario-data" flywheel to support evolution.
- Create "super-individuals", achieve ultra-high productivity, and support sub-scenarios.
- Generate profits from the start, support agile iteration and continuous evolution, and ensure rapid growth.

Navigating uncertainties

Key: Provide an inclusive environment for AI-native.

Challenges faced by the AI-native industry as it moves from concept to reality

High expectation vs. Low ROI

Expectations not met
Pricing dilemma

Model-centric fallacy vs. Engineering reality

The 80-20 rule of AI implementation
A lack of engineering capabilities

Data-driven vs. Insufficient data

Low level of data assetization
Data silos and barriers for data flow

Technological innovation vs. Security & ethics

Unpredictability of agents
Urgent need to align autonomous AI with human values

AIIA works to facilitate the development of the AI-native ecosystem

01 Develop a standards system

02 Build an evaluation and assessment system and develop related capabilities

03 Facilitate industry-wide collaboration

Looking ahead in 2026 and beyond

Drive new breakthroughs: Enhance technical capabilities

- **Continued expansion:** Enhanced hardware-software synergy + New model architectures
- **Continuous learning:** Long-term memory facilitates continuous learning.
- **World models:** Common requirements for the evolution of agents and embodied AI
- **The era of experience:** Reinforcement learning becomes increasingly important during interaction.

Deepen roots: Ensure application

- **Agile iteration:** Adapt to the new normal and obtain value amidst rapid changes.
- **Flywheel effect:** Manage scenarios, application, data, and results in a closed loop.
- **Long-term planning:** Create systematic paths for enterprises to go intelligent.
- **Ecosystem shaping:** Build a new intelligent computing ecosystem characterized by open source and open access.

CAICT 中国信通院

Thank You!

www.caict.ac.cn