CIC灼识

Global Foundation Model Industry Report

Executive Summary

The global foundation model industry has shifted from experimental research to the core engine of global intelligent transformation, using superior scalability and generalization to build a unified intelligence layer for high-order cognitive demands.

Table of Contents

1. Market Overview

1.1 Market Definition

1.2 Market Size and Growth

2. Key Growth Drivers and Trends

2.1 Key Drivers

2.2 Key Trends and Competitive Barriers

2.3 Future Outlook

1. Market Overview

1.1 Market Definition

The global foundation model industry represents a transformative segment within artificial intelligence, functioning as the primary engine for the intelligent metamorphosis of global societies. By unlocking unprecedented levels of productivity and cognitive creativity, these models are redefining the boundaries of human potential. Distinct from traditional small-scale AI models confined to fragmented scenarios, foundation models are engineered with intrinsic scalability and superior generalization.

Foundation model technology companies, the core innovators of the industry, are further categorized into two types:

Pureplay companies: Entities whose core resources, technological accumulation, and business models are entirely centered on the research, development, and commercialization of foundation models, driving rapid technological innovation through focused resource investment.

Non-pureplay companies:

Large internet platforms and cloud service providers leverage capital

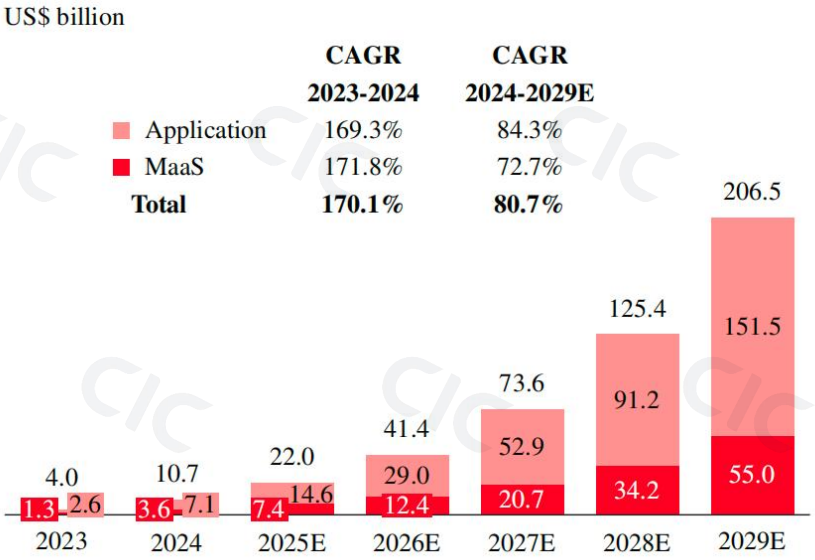
and computing power to integrate foundation model technology into their product ecosystems, accelerating technical validation and commercialization.

1.2 Market Size and Growth

Revenue in the global foundation model market stems from two approaches: model-based and deployment-based, with the former serving as the primary growth engine. Model-based revenue is derived from AI-native applications via subscription models and Model-as-a-Service (MaaS) through cloud-based APIs and licensing, while deployment-based revenue focuses on customized on-premise solutions.

Driven by maturing technologies and increasing user willingness to pay, the global model-based foundation model market is poised for explosive growth. According to CIC, this market is projected to expand from US$10.7 billion in 2024 to US$206.5 billion by 2029, representing a CAGR of 80.7%. Within this segment, the application market is expected to reach US$151.5 billion, while the MaaS market will grow to US$55.0 billion by 2029.

The global foundation model market size, in terms of model-based revenue, 2023-2029E



Source: CIC Reports

Note: Model-based revenues primarily include income generated from foundation model application subscriptions, and foundation model API calls and licensing.

2. Key Growth Drivers and Trends

2.1 Key Drivers

Technological Leaps

The global foundation model industry is defined by disruptive technological breakthroughs, where each generation of model iteration unlocks unprecedented application boundaries and commercial value. Notably, innovations such as the "interleaved thinking" framework and enhanced coding capabilities in Claude 3.7 and Claude Code have shifted the industry paradigm from passive response tools to active AI agents capable of autonomous task orchestration.

Scaling Law

The scaling law remains a fundamental driver underpinning industry growth. The pre-training scaling law—where performance improves with model scale, data and compute—still applies to text, audio and video.

A new test-time compute scaling law has emerged: as shown in 2025 top reasoning models, greater inference compute enhances intelligence. The synergy between pre-training and inference scaling

is forming a new "Moore's Law" for the industry, driving collective progress in model throughput and complex problem-solving capabilities.

Cost Reduction and Market Deployment

A more predictable driver than capability enhancement is the continuous decline in inference costs, which have plummeted from approximately US$20 per million tokens in late 2022 to below US$0.1 by late 2024.

Driven by architectural innovations, engineering optimizations, and falling computing costs, this expected tenfold annual decline makes previously unviable vertical applications, such as large-scale content moderation and real-time AI companionship, commercially feasible. This cost-efficiency trend significantly lowers adoption barriers, accelerating the widespread integration of foundation models into high-volume industrial and consumer scenarios.

2.2 Key Trends and Competitive Barriers

Sustained Improvement in Model Intelligence

Expansion of model scale and capabilities

Foundation models have witnessed a sharp rise in parameters and

notable performance gains, with GPT series models showing near-human reasoning and comprehension abilities in professional assessments. The Mixture-of-Experts (MoE) architecture has become a key breakthrough, expanding model scale while controlling computational costs and latency.

Improving context windows and reasoning efficiency

Modern foundation models have transitioned from the 2,048-token limit of GPT-3 to multi-million token context windows, enabling high-fidelity interaction with ultra-long documents. However, longer context windows raised inference costs, prompting architectural innovations and retrieval-augmented generation. Among the most notable innovations are improvements to the attention mechanism.

Alignment with humans

RLHF (reinforcement learning from human feedback) has become a standard training procedure for foundation models, improving their adherence to user instructions and response quality. Human-aligned models have achieved marked improvements in accuracy, tone control, and handling of inappropriate queries.

Emergence of CoT (chain-of-thought) and reasoning models

The CoT (chain-of-thought) prompting technique, introduced in 2022, improved performance on complex reasoning tasks. A major shift in 2024 saw models trained to break down problems step-by-step during inference, allocating more compute to iterative reasoning, reflection, and output refinement. Reasoning is a computable process, not just emergent from model size. The cost-latency-quality trade-off will split future models: one for fast, accurate responses, the other for deep, resource-intensive reasoning, with test-time compute as key.

Agentic tool use as a new paradigm

AI agents have become a new development paradigm, enabling models to autonomously plan and use external tools for complex tasks. In 2023, GPT-4's plug-ins and function calling broke the limits of training-data-only operation, and Gemini realized autonomous code running in a sandbox. In 2025, leading companies enhanced models' agentic capabilities, turning them from passive responders into active task orchestrators.

Parallel development of closed- and open-source models

Closed-source and open-source models have advanced in parallel

in recent years. The open-source trend has pushed closed-source developers to iterate faster and provided users with more customizable options.

Acceleration of progress

The intelligence level of foundation models worldwide continues to advance. According to OpenAI's five-level roadmap, current models have now reached the threshold of Level 3. Looking ahead, the trajectory points clearly towards accelerated progress.

Levels	Name	Description
L1	Chatbots	AI with conversational language
L2	Reasoners	AI with human-level problem solving
L3	Agents	AI that can take actions
L4	Innovators	AI that can aid in invention
L5	Organizations	AI that can do the work of an organization

Source: OpenAI

Continuous Expansion of Modalities

From single-modal to multi-modal

Foundation models have expanded into the multi-modal domain, aiming to integrate and align features from text, image, audio, and video into a shared semantic space, enabling integration across

different modalities.

Visual understanding

In the early stages of multi-modal understanding, but recently, the trend has been shifting towards more unified multi-modal capabilities. GPT-4V, for example, extends the GPT-4 framework to support image inputs, allowing users to ask the model to analyze visual content, describe image details, interpret humor in memes and information in medical images. Built on a decoder-only architecture, Gemini supports image, video, and audio modalities, with its new benchmarks in multi-modal reasoning tasks.

Audio generation

Text-audio fusion enables AI to interpret and generate audio. Audio synthesis advanced rapidly in 2023: ElevenLabs and MiniMax produced human-like voices, while OpenAI added real-time voice conversation to ChatGPT, boosting applications in smart assistants and customer service.

This integration spawned voice AI agents and smart devices. Future models will better grasp speech emotion and intent, delivering more natural responses and smoother human-machine interaction.

Visual generation

By 2022, the convergence of diffusion models and Transformer-based architectures enabled the generation of high-fidelity, artistic imagery from textual prompts. Since 2023, this momentum has shifted toward video generation, streamlining content production and enhancing creative efficiency across industries. Current research is pioneering unified multi-modal models that fuse the autoregressive reasoning of LLMs with the generative precision of diffusion, mirroring the integrative nature of human intelligence.

Rising Adoption of Foundation Model Applications Unlocking Commercial Value

Unprecedented growth of foundation model applications

Over the past three years, next-generation AI has seen hyper growth, outpacing all previous technological waves in history. Related AI products have achieved extremely fast user scale and commercial revenue growth. Supported by existing technological infrastructure, AI has been spreading rapidly across the internet. Humanity is at a pivotal inflection point of exponential technological growth. While

progress may seem linear in real time, exponential leaps often take place within a short period.

Massive TAM of foundation model applications driven by generalization

Foundation models are revolutionizing productivity, entertainment, and 2B services by enabling both mass deployment and personalized applications through a single, scalable architecture. This versatility delivers high model ROI across a user spectrum ranging from global enterprises to individual creators.

2.3 Future Outlook

The commercialization of foundation model applications is still in the early stage, and agent integration acts as the critical inflection point poised to unlock substantial commercial value.

Agent Applications: AI agents are evolving from tool providers into end-to-end value deliverers, shifting the market focus from enterprise software to the multi-trillion-dollar global labor services market.

Entertainment and Generative Applications: Driven by demand for immersive co-creation, personalized AI companions with integrated

emotional intelligence are redefining the entertainment landscape and fostering deeper user bonds. This shift is amplified by the democratization of content creation via AI video generation and the emergence of advanced voice interfaces as the standard for multi-modal interaction, collectively streamlining human-machine communication and accelerating growth across diverse vertical sectors.

Multi-modal Applications: In March 2025, GPT-4o's native multi-modal upgrades significantly improved image quality and subscriber growth, validating the commercial potential of integrated multi-modality. By enabling precise in-image text and fine-grained editing, the model unlocked professional applications in education, marketing, and scientific illustration. Future deep integration of text, audio, and vision is poised to realize fully editable, synchronized video creation, driving the next global revolution in short-form content.

Notably, driven by the synergy of scaling laws and plummeting inference costs, foundation models are no longer a distant frontier. As AI penetrates vertical industries like finance and manufacturing, competition will increasingly hinge on three core barriers: frontier R&D capabilities, commercialization efficiency, and talent-centric organizational strength, defining the next generation of industry leaders.

In the future, the global foundation model industry is converging into a mature, agent-driven ecosystem that will serve as the indispensable core infrastructure for human society, fundamentally reshaping global production and solidifying its role as the primary engine of a new productivity revolution.

About CIC

CIC is a professional consulting firm offering tailored end-to-end support across the full investment and financing lifecycle. The firm boasts a world-leading track record in guiding landmark first-in-sector IPOs across global markets, alongside unrivaled reach and in-depth coverage capabilities across specialized niche market segments.

CIC helps enterprises refine scalable business models and craft compelling capital narratives to enable seamless access to global capital markets, while serving as a trusted due diligence partner to investment institutions. It delivers granular industry insights and direct access to subject matter experts, empowering clients to identify high-value opportunities and mitigate critical risks effectively.

CIC team maintains deep, real-time market intelligence across a diverse set of sectors—including financial services, artificial intelligence, big data, internet, high technology, healthcare, education, entertainment, consumer goods, transportation and logistics, energy and power, environmental and building technology,

chemicals, industrial manufacturing, and agriculture—delivering unparalleled access to sector-specific, actionable insights.

CIC Reports & Industry Overview

At CIC, we employ a rigorous, multi-method research framework, combining primary and secondary sources to underpin our analysis. Primary research involves in-depth engagements with industry thought leaders and practitioners, particularly in supply chain finance. Secondary research synthesizes publicly available datasets from authoritative bodies, including the National Bureau of Statistics of the People's Republic of China, the State Administration of Financial Regulation (SAFR, formerly the China Banking and Insurance Regulatory Commission), the China Securities Regulatory Commission (CSRC), and public company filings. We apply proprietary data analytics frameworks to process collected information, validating findings through cross-referencing data from multiple research streams to ensure analytical rigor and reliability.

All statistical data presented is verifiable and grounded in information available as of the date of this report.

Extracts are refined summaries of in-depth CIC industry research reports, highlighting supply and demand trends, key growth drivers, R&D trends and future outlook, etc. of various segmented fields, integrating multi-dimensional insights such as expert interviews, market surveys and industry data analysis.

Disclaimer

This report (the "Report") is prepared by CIC based on information available as of the date hereof. The Report is furnished strictly for informational and reference purposes only and is not intended to, nor shall it be construed as, being definitive or conclusive. Nothing contained herein shall constitute or be deemed to constitute investment advice, a recommendation, or an offer, solicitation or inducement to engage in any investment activity. CIC hereby expressly disclaims any and all liabilities for any loss, damage or claims of any nature howsoever arising, whether directly or indirectly, from the use of or reliance upon any information contained in the Report.

Contact CIC

For more information about this report or to learn more about CIC services, please visit CIC official website, or email us at marketing@cninsights.com.