The information and statistics set out in this section and other sections of this document were extracted from different official government publications, available sources from public market research and other sources from independent suppliers, and from the independent industry report prepared by Frost & Sullivan in connection with the [**REDACTED**] *(the "Frost & Sullivan Report"). The information from official government sources has not been independently verified by us, the Joint Sponsors,* [**REDACTED**]*, any of their respective directors and advisors, or any other persons or parties involved in the* [**REDACTED**]*, and no representation is given as to its accuracy.*

SOURCE OF INFORMATION

We engaged Frost & Sullivan, an independent market research consultant, to conduct an analysis of, and to prepare the Frost & Sullivan Report about, the token supply platform market. Frost & Sullivan is an independent global consulting firm founded in 1961 in New York and its services include, among others, industry consulting, market strategic consulting and corporate training. In connection with the market research services provided, we have paid a fee of RMB550,000 to Frost & Sullivan, which we believe to be consistent with market rates. In compiling and preparing the Frost & Sullivan Report, Frost & Sullivan conducted (1) primary research includes interviewing industry participants, competitors, downstream customers and recognized third-party industry associations; and (2) secondary research includes reviewing corporate annual reports, databases of relevant official authorities, as well as the exclusive database established by Frost & Sullivan over the past decades. The market projections in the Frost & Sullivan Report are based on the following key assumptions during the forecast period: (i) the social, economic and political conditions in Chinese markets discussed will remain stable during the forecast period, (ii) government policies on China's token supply market will remain consistent during the forecast period, and (iii) token supply market will be driven by the factors which are stated in the Frost & Sullivan Report. All the data and forecasts contained in this section are derived from the Frost & Sullivan Report. The commissioned report has been prepared by Frost & Sullivan independently without influence from the Company or other interested parties. Our Directors confirm that, to the best of their knowledge, after making reasonable inquiries, there is no material and adverse change in the market information since the date of the Frost & Sullivan Report, which may qualify, contradict or have an impact on the information in this section.

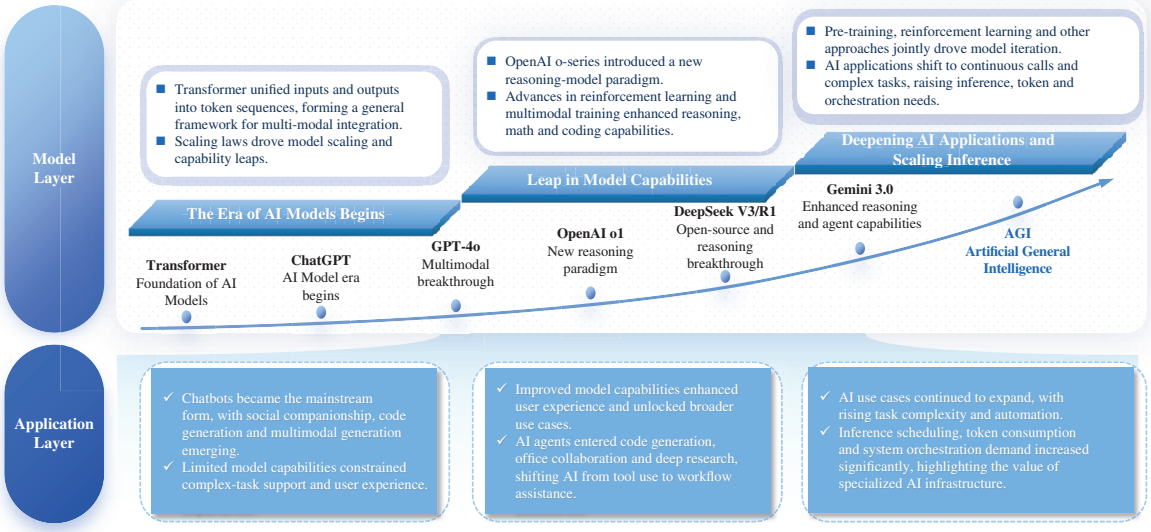
VALUE ANALYSIS OF THE AI INFRASTRUCTURE LAYER AGAINST THE BACKDROP OF THE FOURTH INDUSTRIAL REVOLUTION

AI is becoming the core driver of a new round of industrial revolution and productivity revolution. Following the eras of steam power, electricity and information technology, AI is driving the global economy and society into a new stage of development characterized by intelligence. In terms of market size, as measured by revenue, the global AI market increased from approximately RMB2.2 trillion in 2021 to approximately RMB5.9 trillion in 2025, representing a CAGR of 27.6%, and is expected to further increase to approximately RMB26.3 trillion in 2030 at a CAGR of 36.0%.

Compared with previous industrial revolutions, AI features faster technological iteration, broader application diffusion and stronger penetration into and empowerment of various industries. It not only improves efficiency, but also systematically reshapes productivity systems, business models and company organizational structures. Each industrial revolution has been underpinned by a set of infrastructure: railways in the steam age, power grids in the electricity age, and communications networks in the information age. The same applies to the current wave — AI infrastructure, which supports model training, inference, resource orchestration and service delivery, serves as the fundamental pillar enabling AI's value realization and scalable expansion.

Development History and Characteristics of the AI Industry

Development History of Global AI Industry



Source: Frost & Sullivan

- ***The AI industry remains in its early-to-mid stage, with computing resources and AI models benefiting first; in the long term, as applications are gradually commercialized, the value of system software will continue to boom***

The AI industry primarily consists of the application layer and the infrastructure layer. The application layer serves massive users and a wide range of vertical industry scenarios by providing various AI agents, software/hardware products and intelligent solutions. The infrastructure layer mainly comprises computing resources, AI models and system software.

The AI industry is currently in its early-to-mid stage. Computing resources and AI models, as the key foundation for the generation, deployment and invocation of AI capabilities, have higher certainty and have therefore gained earlier commercial and capital market recognition. In the long term, as AI applications accelerate toward large-scale commercialization, the value of upper-layer applications will gradually become more prominent, creating larger-scale commercial opportunities. In this process, the value of the system software will continue to boom with the growth in the number of applications, invocation frequency, model complexity and enterprise-grade deployment needs. On the one hand, it continuously improves the utilization efficiency of heterogeneous computing resources and various AI models; on the other hand, it supports application-side requirements for high concurrency, low latency, stable delivery and overall cost optimization, thereby becoming a key layer connecting the upper and lower ends of the AI industry stack.

- ***The large-scale deployment of AI applications and value realization are accelerating the migration of computing demand from training to inference***

Training and inference are the two main stages through which AI capabilities are formed and their value is realized. Training determines the foundation of model capabilities and is typically a centralized investment during the capability formation stage. Inference, by contrast, takes place during the actual use of AI models and represents the value realization stage that determines whether model capabilities can be continuously delivered to real-world users, real-world scenarios and real business operations. It is characterized by high frequency, continuity, distribution and real-time requirements.

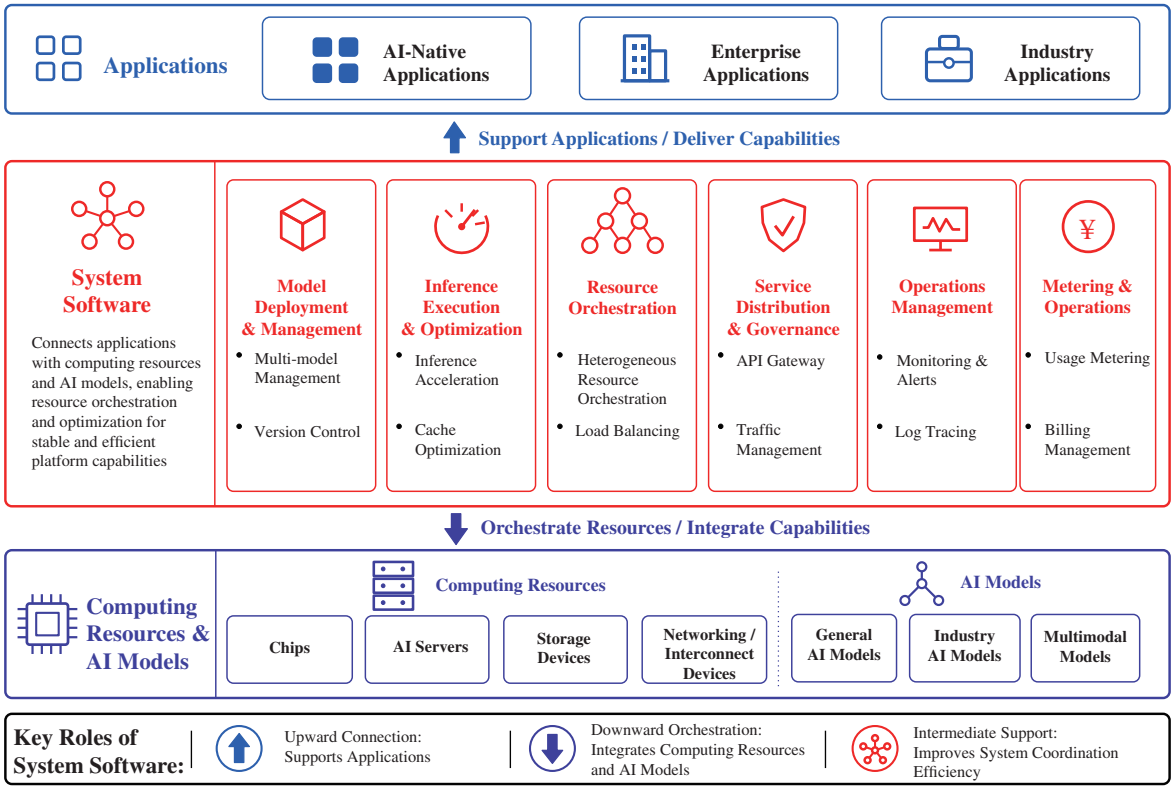
As demand from real-world scenarios surges, AI chatbots, code generation and software engineering assistants, enterprise knowledge bases, text-to-image/video creation tools, autonomous driving solutions and various intelligent applications are continuing to mature. AI applications are evolving from basic scenarios such as single-turn Q&A and text generation toward multimodal interaction, complex task execution and continuous intelligent services. This has led to higher-frequency user interactions, longer context processing, stronger real-time response requirements and increasingly complex tasks, significantly driving up inference demand and the scale of token throughput for AI models. As a result, the center of computing resources in the AI industry is accelerating its shift from training to inference, and the total volume of tokens consumed by a single model over its lifecycle will far exceed that consumed during its training stage.

Analysis of the Elements of the AI Infrastructure Layer and the Key Role of System Software

- ***The AI infrastructure layer mainly comprises computing resources, AI models and system software***

The development of AI applications relies on the continuous enhancement of infrastructure capabilities. The AI infrastructure layer mainly comprises computing resources, AI models and system software. Computing resources include AI chips, AI servers, storage devices and network/interconnect equipment, providing the underlying resource base and operating environment for model training and inference. AI models include general-purpose AI models, industry-specific AI models and multimodal AI models, determining the boundaries of AI capabilities and the scope of applications. System software serves as the intermediary layer for organizing and utilizing computing resources, undertaking functions such as model deployment, inference execution, heterogeneous computing resource orchestration, service distribution, thereby maximizing the potential of computing resources and AI models.

Key Role of System Software



Source: Frost & Sullivan

- ***Token supply platform has emerged as an important product and commercial form of AI infrastructure system software***

By adapting to and encapsulating computing resources and AI models, system software provides downstream applications with callable, manageable and continuously operable AI capabilities. Against the backdrop of the large-scale expansion of AI applications and the rapid growth of inference demand, the token supply platform has emerged in the industry as a product and commercial form built upon AI infrastructure system software. A token supply platform integrates key system software capabilities such as inference engines, heterogeneous computing resource orchestration, model deployment, service distribution, operations management, metering and token delivery. By organizing and orchestrating these capabilities around token production, orchestration, optimization and delivery, the token supply platform transforms dispersed computing resources and models into stable, efficient and scalable token services.

- ***Token supply platform emerges as a key value-creating segment for cross-chip adaptation, model deployment and inference efficiency improvement***

Currently, AI inference is characterized by the coexistence of multiple elements across chip architectures, computing resources, model types, deployment methods and customer needs. Meanwhile, the continuous launch and iteration of new chips and AI models further raises the requirements for rapid adaptation, migration and performance optimization at the inference level. On the computing resource side, different chips vary in hardware architecture, software ecosystem and maturity of adaptation. On the model side, different models have their own focuses in terms of parameter scale, context length, inference efficiency, and applications, with AI models, as well as general-purpose and vertical models, iterating rapidly. On the customer side, enterprises have developed differentiated demands for serverless token services, dedicated instances, on-premise deployment.

This diversified landscape has promoted specialization across the industry chain, while also significantly increasing the complexity of large-scale deployment of "cross-chip × multi-model" combinations across different types of customers and application scenarios. Therefore, the core value of the system software layer as a token supply platform has become increasingly prominent. On the one hand, it adapts to and encapsulates various computing chips and models, and conducts in-depth optimization of operators, precision, chip memory and inference frameworks, thereby ensuring the delivery of stable inference services across different applications. On the other hand, through inference engine optimization, computing resource orchestration, and service and operations management, it continuously improves the token production efficiency per unit of computing resource on existing hardware, while ensuring stable service and scalable delivery.

Therefore, as AI applications expand at scale, demand for AI infrastructure system software oriented toward inference scenarios will continue to grow. Token supply platform with cross-chip, multi-model and multi-scenario adaptation capabilities will become a key infrastructure supporting the large-scale deployment and value realization of AI industry applications.

The Token Revolution and Its Far-reaching Impact Amid the Rise of AI Applications

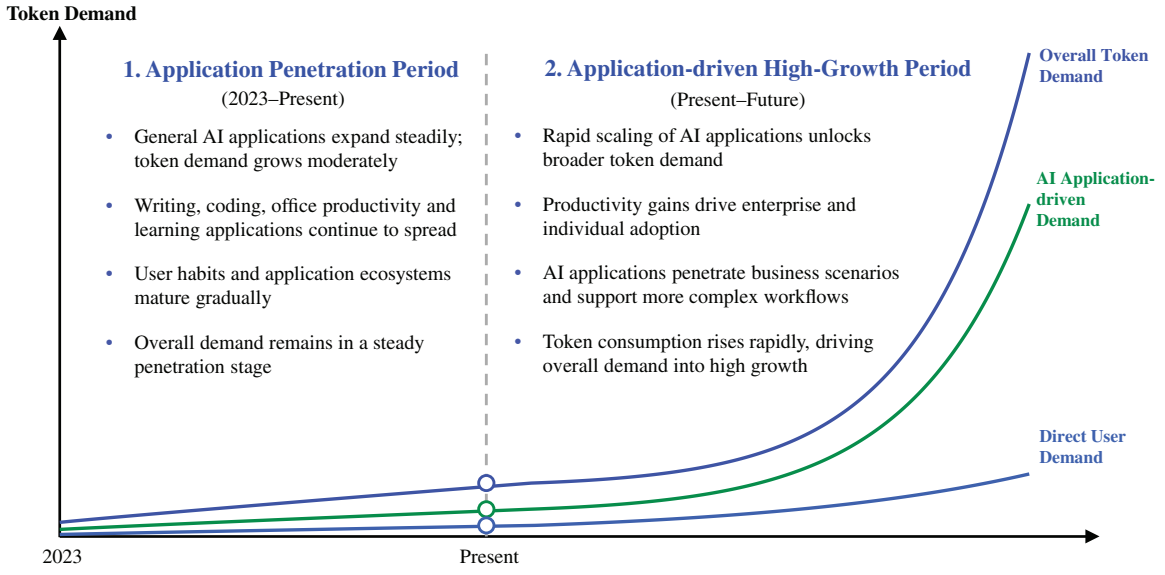
- ***Tokens are evolving from a unit of measurement for model usage into a standardized product form for end applications***

AI applications are evolving from single-turn Q&A and text generation toward multimodal interaction, complex task execution and continuous intelligent services, giving rise to more frequent and continuous demand for token service. In 2025, China's token throughput expanded rapidly to approximately 2,426.3 trillion tokens. Tokens are no longer merely a unit of measurement in the input and output process of AI models, but are gradually becoming a standardized product form for end applications. Token supply platform produces and delivers tokens through the capabilities of computing resources, AI models and system software, while AI applications deliver intelligent services through the continuous consumption of tokens.

- ***Token quality and production efficiency determine effective supply capacity and become key to supporting the large-scale deployment of AI applications***

As the large-scale deployment of AI applications unlocks broader token demand, the key focus of industry competition is no longer merely the scale of underlying computing resources, but whether dispersed and heterogeneous chip resources, model capabilities and deployment environments can be efficiently converted into stable and usable token production. While token demand is growing rapidly, the supply side remains subject to physical constraints in terms of chips, memory capacity, electricity, data centers and computing cluster construction, as well as supply chains, construction cycles and resource allocation. Under the dual catalysis of rapidly released demand and relatively limited supply, improving token production efficiency and enhancing the stability and availability of token services have become key capabilities supporting AI applications in moving from pilot exploration to large-scale deployment.

Evolution of Token Demand Driven by the Rise of AI Applications



Source: Frost & Sullivan

GLOBAL AND CHINA TOKEN SUPPLY MARKET ANALYSIS

Current Status and Pain Points of the Token Supply Industry

- ***Heterogeneous Computing Environments Increase Adaptation and Orchestration Complexity, Constraining Token Quality and Production Efficiency***

From the perspective of the AI chip ecosystem, global computing resource supply is primarily supported by NVIDIA and domestic vendors. Different chips vary significantly in hardware architecture, software ecosystems, performance boundaries, adaptation maturity and cost structures. On the one hand, such differences raise the barriers to cross-chip deployment adaptation, operator optimization and performance tuning for the same model. On the other hand, cross-chip hybrid clusters further increase the complexity of overall computing resource orchestration, making it difficult for heterogeneous computing pools to orchestrate inference workloads evenly and often resulting in imbalanced hardware utilization and idle resources. If enterprises lack mature capabilities in cross-chip adaptation, deep inference engine optimization and end-to-end engineering tuning, key metrics such as token throughput, cluster resource utilization, efficiency, user experience and cost per-token may fluctuate significantly. As a result, the underlying computing resources may fail to fully unlock peak performance or reliably support massive AI inference workloads, thereby increasing the overall cost of token production and substantially weakening the economic viability of large-scale token service procurement by enterprise customers. Accordingly, competition within the AI infrastructure layer is shifting from the simple accumulation of computing resources toward token production efficiency.

- ***Highly Diverse Models and Deployment Ecosystems Require Stable Token Services***

From the perspective of the model ecosystem, multiple AI models, including DeepSeek, Qwen, GLM and Kimi, currently coexist in the market. These models differ in parameter scale, architecture, inference efficiency, context length, deployment requirements and applications. Together with enterprises' diversified requirements for deployment approaches, data security and model selection, the industry has gradually evolved into an inference infrastructure landscape characterized by the coexistence of cross-chip, multi-model and multi-deployment approaches. For enterprise customers, the value of token services depends not only on model capabilities themselves, but also on whether tokens can be invoked in a stable, continuous and low-latency manner while maintaining a stable experience across different business scenarios. This requires an intermediate infrastructure layer to centrally manage, deploy and orchestrate heterogeneous computing resources and AI models, thereby ensuring token production speed, service stability and scenario adaptability.

The Role of Token Supply Platform Across the Token Production Lifecycle

In an industry environment characterized by the coexistence of diverse computing resources, multiple models, and multiple deployment approaches, token supply platform serves as a system software layer connecting computing resources, models, and applications. By organizing, orchestrating, and optimizing the entire token production and delivery process, the token supply platform addresses key industry challenges, including fluctuations in token production efficiency, uneven resource utilization, and rising requirements for service stability.

- ***Token Production Lifecycle: Underlying Resource Access, Application Demand Intake, Model Selection, Heterogeneous Computing Orchestration, Inference Execution, Token Encapsulation, and Result Delivery.***

Token production is a continuous service process comprising computing resource orchestration, model deployment, and application delivery. First, token production requires access to infrastructure, including energy, data, models, computing clusters, high-speed interconnects, and liquid cooling systems, to form the underlying resource base available for invocation. Subsequently, users initiate token requests via APIs, application interfaces, agents, or various application systems. The platform then performs model selection, task routing, and resource matching based on request content, task type, model capability, context length, response latency, SLA requirements, etc.

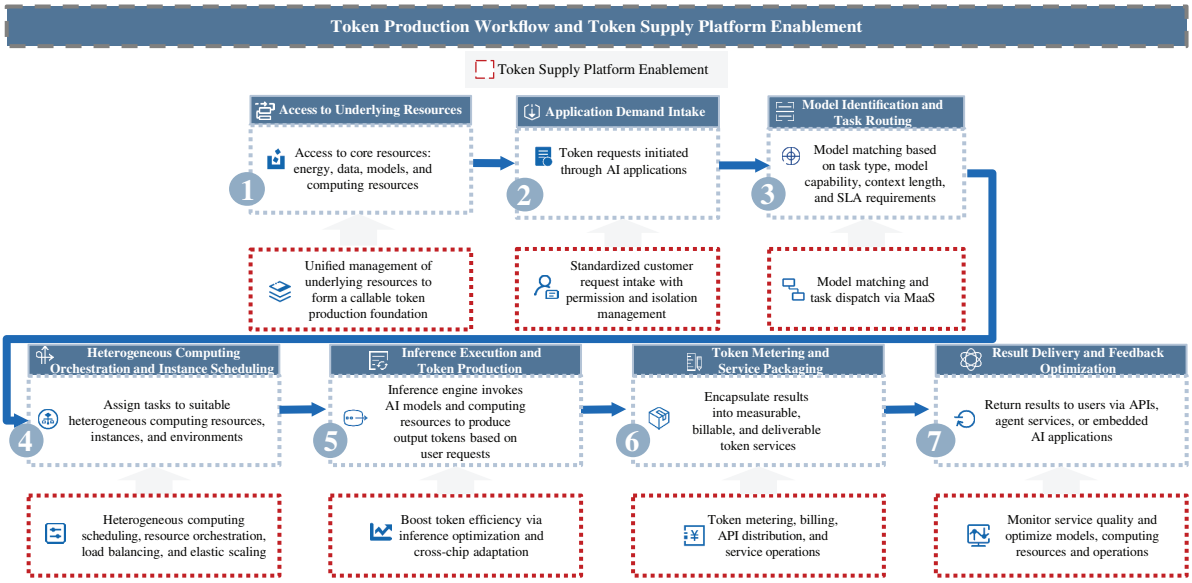
Once inference begins, the platform allocates tasks to appropriate heterogeneous computing resources, inference instances, and runtime environments. The inference engine invokes underlying chips, servers, and models to perform computation and produce output tokens based on user requests. Finally, the generated results are encapsulated into standardized token services that are measurable, billable, and deliverable. They are delivered to users via various AI applications. Meanwhile, the platform continuously monitors service quality, resource utilization, and unit token cost, using these insights to optimize model deployment, computing orchestration, and operational strategies.

- ***Token Supply Platform Empowers the Entire Token Production Lifecycle: Delivering Strategic Value Through Enhanced Token Supply Efficiency, Stability, and Scalability***

In the token production value chain, token supply platform serves as the central layer connecting computing resources, models and applications. Through capabilities such as cross-chip adaptation, model deployment, inference optimization, heterogeneous computing resource orchestration, elastic scaling, service distribution, metering and billing, and service monitoring, token supply platform empowers the entire token production and delivery lifecycle. By transforming complex inference demands into standardized, manageable, and scalable token services, token supply platform improves supply efficiency, reduces the complexity of cross-chip and multi-model adaptation, and enhances service stability and continuity.

For customers, token supply platform provides not merely computing resource or model access, but encapsulated, ready-to-use token supply services featuring cost efficiency, stable performance and multi-scenario adaptability. For the computing and model ecosystem, they enable heterogeneous computing resources and multi-model capabilities to be more efficiently transformed into token production capacity through engineering optimization and service-oriented delivery. As such, token supply platform has become a critical infrastructure layer supporting the large-scale deployment of AI applications.

Token Production Workflow and Token Supply Platform Enablement



Source: Frost & Sullivan

Market Drivers Analysis

- ***Demand: Diverse Model Invocation Needs Driven by Differentiated Model Capabilities***

Different models exhibit distinct strengths, capabilities, and application scenarios, making it difficult for a single model to meet all business needs. As a result, the same enterprise, business unit or even single product often needs to invoke multiple models simultaneously to match different task requirements and scenario-specific needs, driving continued growth in token consumption and increasing demand for highly adaptive multi-model token supply services.

- ***Supply: Cost Reduction and Efficiency Gains Drive the Development of Specialized Token Supply Services***

As enterprise-grade AI applications enter production environments, customers are placing greater emphasis on token access costs, response speed, service stability, and invocation efficiency. Through capabilities such as inference optimization, resource orchestration, caching mechanisms, and service operations, specialized token supply platform can improve token production efficiency, reduce generation and invocation costs, and enable customers to access AI capabilities more efficiently.

- ***Technology: Coexistence of Heterogeneous Computing Resources and Multi-Model Ecosystems Drives Growing Adaptation Needs***

The market is characterized by the coexistence of heterogeneous computing resources, including NVIDIA and various domestic AI chips, as well as a broad model ecosystem comprising DeepSeek, Qwen, Kimi, GLM, and Wan. Industry policies, the development of domestic AI chips, competition among multiple technology pathways, and increasingly differentiated deployment requirements have collectively reinforced a long-term heterogeneous landscape, further increasing the importance of cross-chip adaptation, model deployment, inference optimization, and unified scheduling capabilities.

- ***Ecosystem: Diverging Enterprise-grade Deployment Needs Enhance the Value of Independent Ecosystem***

Enterprise customers are exhibiting increasingly diverse requirements for public/private deployment/operation, resource and access isolation, and scenario-specific adaptation, driving token supply services to evolve from a single service mode toward multiple service modes. Independent ecosystem token supply platform can connect heterogeneous computing resources, diverse model ecosystems, and various demands of enterprise customers, providing more flexible model choices, deployment options, and token service capabilities within a multi-stakeholder ecosystem.

Main Industry Service Modes

Token supply platform service modes are typically linked to customers' token consumption volumes, resource utilization patterns, service performance requirements, and deployment methods. Based on different service forms, the industry has primarily developed three service modes: serverless token services, dedicated instances, and on-premise deployment solutions.

Among these service modes, serverless token services provide token services on a shared-resource basis and are generally charged based on usage volume measured by tokens consumed, primarily targeting developers, small- and medium-sized enterprises and other customers seeking flexible and cost-effective access to AI capabilities. Dedicated instances provide reserved or dedicated computing capacity on public cloud infrastructure, with pricing based on reserved computing capacity and underlying infrastructure costs, and are designed for customers with higher requirements for performance or stability. On-premise deployment solutions deploy our inference engine and computing resource orchestration system within the customer's own data center or private computing environment, enabling customers to use tokens across a wide range of models on their own computing resources, and are primarily targeted at large enterprise customers and institutional clients.

Main Industry Service Modes

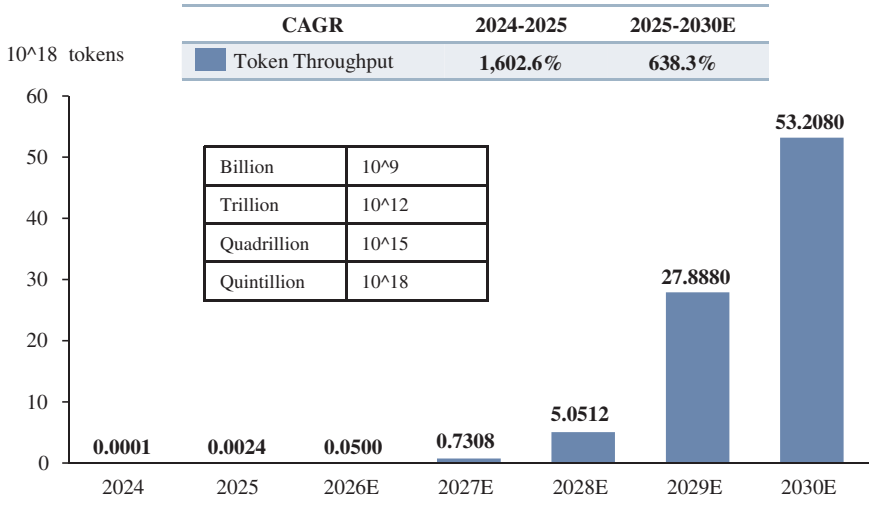
	Serverless Token Services	Dedicated Instances	On-premise Deployment Solutions
Service Mode	Standardized token services on a shared-resource basis, accessible through a single standard interface without upfront dedicated resource commitment	Reserved or dedicated computing capacity on public cloud infrastructure for specific customers	Deployment of inference engine and computing resource orchestration system within customers' own data centers or private computing environments
Billing Basis	Usage volume measured by tokens consumed; typically prepaid credits, with postpaid monthly settlement available for certain repeat customers	Reserved computing capacity, taking into account underlying infrastructure costs, including computing resources and associated operational expenses	Software licensing fees and/or professional implementation or maintenance service fees, depending on deployment scale, system configuration and customer requirements
Service Features	Flexible and cost-effective access, suitable for testing, validation and elastic invocation scenarios	Higher performance, stability, suitable for latency-sensitive and high-throughput AI workloads	Enables customers to use tokens across a wide range of models on their own computing resources; based on a standardized, modular platform architecture, with only limited customization required to accommodate customers' specific deployment environments or operational requirements
Target Customers	Developers, small and medium-sized enterprises and other customers seeking flexible and cost-effective access to AI capabilities	Enterprise users running latency-sensitive, high-throughput AI workloads and other customers with higher requirements for performance and stability	Large enterprise customers and institutional clients

Source: Frost & Sullivan

Token Supply Market Size

In 2024, China's token throughput reached approximately 142.5 trillion tokens processed and surged to approximately 2,426.3 trillion tokens processed in 2025, representing a year-on-year growth rate of 1,602.6% from 2024 to 2025. This growth was primarily driven by the continuous enhancement of AI model capabilities, as AI usage patterns have gradually expanded from single-turn Q&A and content generation to sequential task execution, complex goal decomposition, multi-step tool invocation and multimodal interaction. As a result, invocation frequency, context length, concurrency requirements and task complexity have all increased in tandem. Entering 2026, token throughput is projected to further rise to approximately 50 quadrillion tokens processed. By 2030, China's token throughput is expected to reach approximately 53.2 quintillion tokens processed, representing a CAGR of approximately 638.3% from 2025 to 2030. Against this backdrop, the token supply market is poised to unlock long-term, high-certainty market growth potential.

Token Supply Market Size (by Token Throughput in Quintillion, China, 2024-2030E



Source: Frost & Sullivan

Analysis of Market Development Trends

- ***Computing Resource Supply: Persistent Diversity in Computing Resources***

The future supply of computing resources will maintain a diversified landscape characterized by the coexistence of international and domestic AI chips, multiple vendors and parallel hardware architectures, such as GPU and NPU. As computing resource heterogeneity and ecosystem adaptation continue to evolve, multi-source, multi-vendor and multi-architecture heterogeneous orchestration capabilities will become a core industry focus.

- ***Service Mode: Expansion from Single Serverless Token Services to Multiple Service Modes***

As enterprise customers evolve from testing to production-grade deployment, their requirements for token throughput, response stability, latency tolerance, and resource/access isolation will become increasingly differentiated. Token supply platform will expand their service modes from shared-pool API throughput to a parallel mix of API throughput, dedicated instances, and on-premise deployment, satisfying elastic scaling, stable supply, and private deployment needs.

- ***Product Capability: Deepening Focus on Inference Efficiency and Service Stability***

In an environment where multiple models, chips, and deployment approaches coexist, token supply platform needs to enhance technical capabilities across the entire token production lifecycle. The inference engine layer will improve token output efficiency; the cluster scheduling layer will optimize computing resource utilization; and the MaaS and API distribution/operations layer will strengthen multi-model integration, unified invocation, service monitoring, and customer operations — collectively improving the efficiency, stability, and scalability of token services.

- ***Ecosystem: The Growing Connectivity Role of Independent Ecosystem Platform***

Cloud service providers, large tech enterprises, and model vendors typically build closed ecosystems leveraging their own resources, with service systems prioritized to fit their own computing resources, models and applications. In contrast, independent ecosystem token supply platform maintains a neutral position, without favoring any single computing resource or model. It connects multiple types of computing resources, AI models, and various industry participants, delivering unique value through cross-chip, multi-model, and multi-deployment service offerings.

Analysis of Key Success Factors

- ***Cross-chip and Multi-model Adaptation with Inference Optimization Capabilities***

In a market with diverse computing resources and multi-model, enterprises need the capability to understand and adapt to different chip architectures, software ecosystems, inference frameworks, and model structures. Through cross-chip adaptation, operator optimization, model deployment optimization, and inference engine optimization, enterprises can fully unlock computing potential, improve token production efficiency, reduce token production costs, and enhance service availability and stability.

- ***Cluster Orchestration, Elastic Service, and Stable Delivery Capabilities***

As AI applications evolve from single to continuous invocations, complex task execution, and production-grade deployment, token services must operate stably under high concurrency, low latency, and cost-controlled conditions. Enterprises need the capability to orchestrate diverse computing resources, uniformly managing and efficiently dispatching third-party data center capacity, internal computing resources, and on-premise deployment resources. Building on this, unified management of GPU servers and heterogeneous computing clusters — encompassing resource orchestration, instance scheduling, elastic scaling, service monitoring, as well as operation and management — enables improved resource utilization and ensures production-grade service delivery.

- ***Neutrality, Open Positioning, and Multi-Ecosystem Integration Capabilities***

Token supply platform needs to connect chip companies, computing cluster providers, model vendors, developers, and enterprise customers, building multi-computing resource, multi-model, multi-deployment, and multi-scenario token supply capabilities. The platform focuses on infrastructure services and does not compete with application vendors or end customers in business. Compared to closed ecosystem players that combine both application development and service provision, the independent ecosystem token supply platform maintains a more neutral positioning, openly collaborating with various industry participants across the value chain.

COMPETITIVE ANALYSIS OF CHINA'S TOKEN SUPPLY MARKET

Overview of Market Competition Landscape

China's token supply market is at a stage of rapid development and formation of the competitive landscape. Industry competition mainly centers on token production efficiency, heterogeneous computing resource orchestration capabilities, multi-model adaptation capabilities, enterprise-grade service capabilities and connectivity capabilities. Currently, market participants may be classified into two types, namely independent ecosystem and closed ecosystem, with different resource endowments, ecosystem positioning and service boundaries.

Comparison of Different Ecosystems in the Token Supply Market

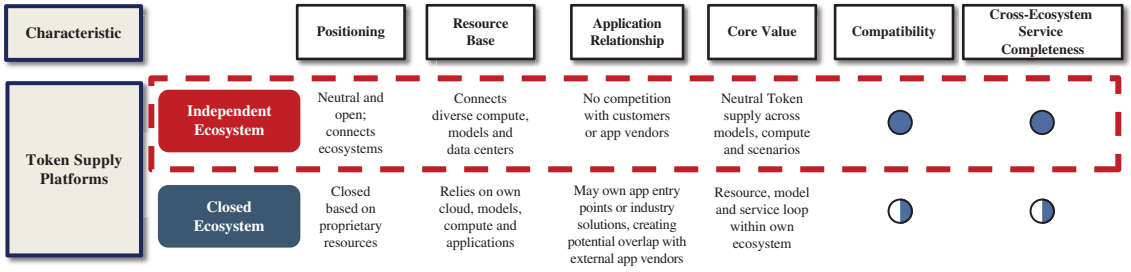
- ***The token supply market has mainly formed two types of business modes: independent ecosystem and closed ecosystem***

Independent ecosystem token supply platform is generally not bound to any cloud provider, model or application scenario. Instead, it provides token supply services across models, chips and deployment environments by connecting multiple types of computing resources, diverse models, different computing centers and enterprise customer needs. Closed ecosystems are typically developed by cloud service providers, large technology companies or model companies based on their own cloud resources, models, computing infrastructure and application scenarios, enabling resource integration, model invocation and service closed-loop within their own ecosystems.

- ***The independent ecosystem has neutral, open and cross-ecosystem connectivity value under a diversified heterogeneous landscape***

Compared with closed ecosystems, independent ecosystem token supply platform features neutrality, openness, diversified connectivity and horizontal adaptation capabilities, allowing customers to freely combine computing resources and AI models based on their needs. Such neutral positioning is more likely to earn customer trust in the multi-party competitive landscape of the AI industry in its early-to-mid stage, and can connect across different ecosystems, thereby accumulating differentiated value that is difficult for closed ecosystems to replace.

Comparison of Participants in the Token Supply Market



Source: Frost & Sullivan

Analysis of Company Ranking and Market Share

Overall Company Ranking and Market Share

In 2025, in terms of annual token throughput, the Company ranked fourth in China's token supply market, with a market share of approximately 1.5%.

Ranking of Token Supply Platforms (by token throughput), China, 2025

Ranking	Company Name	Token Throughput (trillion tokens)	Market Share
			(%)
1.	Company A	1,036.0	42.7
2.	Company B	788.5	32.5
3.	Company C	286.3	11.8
4.	**The Company**	**36.7**	**1.5**
5.	Company D	34.8	1.4
	Others	243.9	10.1
	Total	2,426.3	100.0

Notes:

(1) Company A is a private company established in 2020 and headquartered in Beijing, China. It focuses on cloud and AI services, with services mainly covering cloud infrastructure, video and content distribution, big data, AI, and development and operations.

(2) Company B is a private company established in 2009 and headquartered in Zhejiang, China. It is affiliated with a listed group and provides cloud computing and AI services, with services mainly covering computing, storage, networking, databases, big data, AI model services and cloud-native services.

(3) Company C is a private company established in 2001 and headquartered in Beijing, China. It is affiliated with a listed group and operates AI cloud services, with services mainly covering cloud infrastructure, AI development platforms, model services and industry AI solutions.

(4) Company D is a private company founded in 2019 and headquartered in Shanghai, China. It focuses on distributed cloud computing services, with services mainly covering intelligent computing, model services, edge computing and edge CDN.

Ranking and Market Share of Independent Ecosystem Token Supply Platforms

In 2025, in terms of annual token throughput, the Company ranked first among independent ecosystem token supply platforms in China's token supply market, with a market share of approximately 1.5%.

Analysis of Industry Entry Barriers

- ***Seamless connection and performance optimization barriers across chips and models***

Token supply platforms need to achieve seamless connection across multiple types of chips, diverse models and different inference engines. Relevant capabilities involve multiple aspects, including model deployment, operator optimization, inference engine optimization and multi-computing resource orchestration. Due to significant differences in software ecosystems, operator compatibility and model structures across different chips, such capabilities require long-term engineering practice and continuous iteration and accumulation. New entrants find it difficult to develop reusable and scalable adaptation and optimization capabilities within a short period of time.

- ***Large-scale token service operation barriers***

Token supply platforms need to operate stably under conditions of high concurrency, low latency and controllable costs, which imposes high requirements on enterprises' capabilities in cluster orchestration, resource orchestration, elastic scaling, load balancing, service monitoring and operations management. Especially in production-grade application scenarios, customers have higher requirements for the stability, response speed and continuous delivery capability of token services. Enterprises lacking large-scale operational experience find it difficult to satisfy the long-term needs of enterprise customers.

- ***Ecosystem collaboration and resource acquisition barriers***

Token supply platforms need to connect chip vendors, computing resource providers, model companies, developers and enterprise customers to form service capabilities across computing resources, AI models and scenarios. The accumulation of ecosystem resources depends not only on the number of participants connected to the platform, but also on enterprises' ability to build partnerships and deliver services across diverse partners. New entrants usually need a relatively long period of time to establish partnerships, acquire resources and build customer trust.

- ***Commercial validation and customer loyalty barriers***

The capabilities of a token supply platform need to be continuously validated across multiple service modes, including serverless token services, dedicated instances and on-premise deployment. Through repeated model testing, chip adaptation, deployment migration, inference optimization, computing resource orchestration and service operations, platform providers accumulate reusable engineering know-how and delivery toolchains. As enterprise customers increasingly rely on such platforms for production-level AI inference, stable service performance, accumulated integration experience and higher switching costs help strengthen customer loyalty, enabling experienced platform providers to build sustainable competitive advantages.