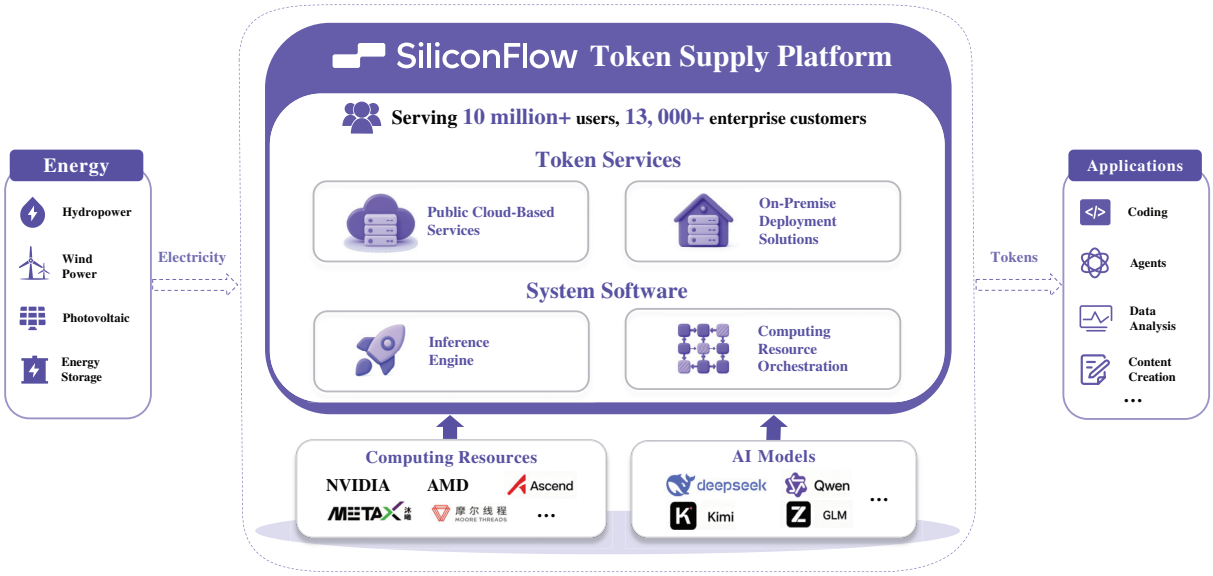
## OVERVIEW

### Who We Are

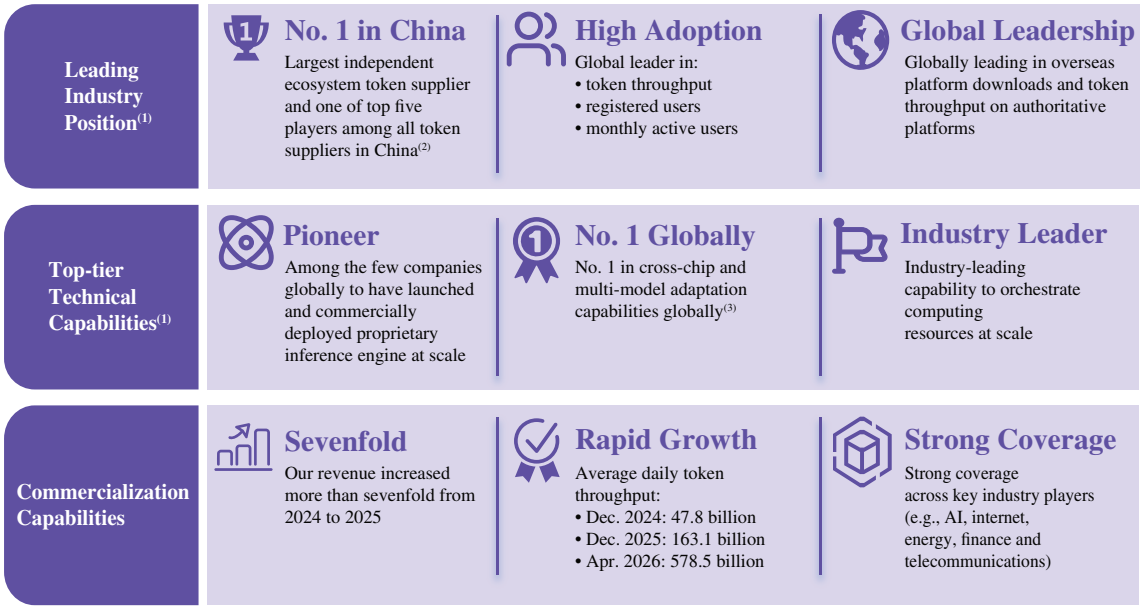
We are a leading open, independent token supply platform in China, operating at the forefront of an era defined by the rapid deployment and application of AI models across enterprises. As AI models become increasingly integrated into everyday workflows, the demand for tokens is growing exponentially, while the need for high-quality token supply, characterized by accuracy, generation speed, cost efficiency, stability and security, is becoming increasingly critical. By connecting computing resources and AI models with system software capabilities, our token supply platform offers one-stop AI infrastructure services, enabling high-quality token delivery at scale.

In China, we are the largest independent ecosystem token supplier and one of the top five players among all token suppliers, as measured by annual token throughput in 2025, according to Frost & Sullivan. Powered by our proprietary inference engine and computing resource orchestration system, we effectively aggregate, optimize, and orchestrate heterogeneous computing resources across diverse hardware architectures from different vendors, thereby converting underlying computing power into standardized, reliable and scalable token supply. This enables seamless access to and deployment of AI capabilities across a broad range of state-of-the-art models with industry-leading performance-to-cost efficiency. To capture diversified commercial demands, we deliver our core capabilities primarily through public cloud-based services and on-premise deployment solutions, serving customers across a wide range of sectors.

The diagram below illustrates our platform in the AI value chain:



As of April 30, 2026, our platform had over 10 million registered users. We recorded average daily token throughput of approximately 578.5 billion and peak daily token throughput of approximately 1,071.4 billion in April 2026. As of the Latest Practicable Date, we had served over 13,000 enterprise customers, and our platform had supported a cumulative total of over 170 models. The following chart highlights some of our key achievements to date:




*Notes:*

(1) According to Frost & Sullivan

(2) As measured by annual token throughput in 2025

(3) Measured by the number of supported chips and models

## Our Value Propositions

The AI technology stack is centered around three core pillars: computing resources, AI models and system software. Computing resources provide the processing power required for AI training and inference; models define the capabilities and intelligence of AI systems, and system software connects computing resources with models and enables AI capabilities to be deployed in real-world applications. As a critical system software pillar within the AI technology stack, we bridge computing resources and models with AI applications, offering a compelling set of value propositions to participants across the AI industry value chain.

***For customers***. We provide customers across a broad range of industries with one-stop access to diverse AI models and infrastructure capabilities, enabling more efficient and cost-effective deployment of AI applications.

- *Simplified deployment.* Our unified and standardized APIs enable customers to rapidly access AI capabilities through a single interface, significantly reducing integration complexity, shortening deployment cycles and lowering technical barriers to adoption.

- *Enhanced performance*. Our proprietary inference engine is designed to improve token throughput, reduce latency and enhance operational stability across different computing environments.

- *Optimized cost efficiency.* Leveraging our cross-chip and multi-model adaptability, we can flexibly match computing resources and models based on customers' specific performance, latency and cost requirements. This enables customers to reduce unit token costs while maintaining service quality.

- *Flexibility and scalability.* Our elastic scaling architecture and diversified deployment options allow customers to tailor computing capacity, AI model selection and deployment architecture based on application needs. This supports the full lifecycle of AI adoption, from research and testing to enterprise-grade production deployment.

***For ecosystem partners***. Our open and independent positioning creates a collaborative environment that aligns incentives and strengthens interoperability across the AI value chain. Through our full-stack software capability – comprising cross-chip and multi-model adaptation, inference engine optimization, heterogeneous computing resource orchestration and standardized token delivery capabilities – we help:

- *chip designers* accelerate the commercialization of AI chips by enabling efficient, scalable and production-ready inference across real-world workloads.

- *computing resource providers* convert fragmented computing resources into commercially deliverable token production capabilities.

- *model developers* achieve scalable and reliable AI inference across diverse computing environments.

- *toolchain and application creators* accelerate product deployment and commercialization.

Our compelling customer value proposition has driven significant expansion of our customer base. As we serve a broader range of customers and use cases, diverse customer requirements support targeted technical optimization, while expanding application scenarios accelerate technology iteration and create additional opportunities for new customer acquisition. As an open and independent platform, we continue to enhance the versatility of our technology by broadening compatibility across a wider range of computing resources and models. By leveraging our accumulated technology capabilities and implementation experience, we are able to reduce adaptation costs and accelerate deployment when entering new computing and model environments. This, in turn, further strengthens our core value proposition and reinforces our influence within the industry ecosystem.

**Our Core Technology Foundation**

We continuously invest in the development of our proprietary inference engine and computing resource orchestration system, both purpose-built for AI inference workloads.

*Full-stack inference engine.* Our inference engine serves as the critical system software operating directly on top of chips. Through technologies such as Expert Parallelism (EP), Prefill-Decode (PD) separation, KV cache optimization, pipeline parallelism, operator optimization, runtime optimization and scheduling optimization, our inference engine transforms underlying computing resources into high-throughput, low-latency and cost-efficient tokens.

*Computing resource orchestration system.* Our computing resource orchestration system is designed for AI inference scenarios characterized by heterogeneous computing environments, high-memory requirements, low-latency inference and bursty high-concurrency demand. Evolving beyond traditional cloud capabilities, our system integrates computing, storage, networking, scheduling, security and observability into a unified framework focused on inference instance scheduling, deployment, scaling, monitoring and operations. Together, our inference engine and computing resource orchestration system form an integrated technology stack supporting large-scale token production.

Leveraging our engineering capabilities across computing resources, AI models and system software, we are able to replicate inference optimization best practices across different models, chip architectures, computing resource providers and customer environments, while rapidly adapting to technology iterations.

## Our Products and Services

We have established a comprehensive software and service product matrix spanning both public cloud and private deployment environments. Through standardized and highly compatible unified APIs and full-stack software capabilities, we provide customers with stable, low-latency and cost-efficient production-grade tokens and full-lifecycle technical support.

*Public Cloud-Based Services*. Our public cloud-based services are designed to deliver standardized, high-quality tokens at scale, providing customers with unified and flexible access to cumulatively more than 170 AI models through a single, standard interface, covering a broad range of AI application scenarios including coding, agents and multimodal applications. We launched our serverless token services in May 2024, marking the transition to large-scale commercialization of our public cloud-based services. The platform supports two complementary service offerings — serverless token services on a shared-resource basis and dedicated instances providing reserved computing capacity.

- *Serverless Token Services*: We provide token services on a shared-resource basis and charge based on usage volume measured by tokens consumed. These services primarily target developers, small- and medium-sized enterprises and other customers seeking flexible and cost-effective access to AI capabilities. We generally adopt a prepaid model for customers, where credits are purchased in advance and deducted based on token usage, and a postpaid model with monthly settlement for certain recurring enterprise customers.

- *Dedicated Instances*: We offer reserved or dedicated computing capacity on public cloud infrastructure, under which computing resources are reserved for specific customers. These services are designed for customers with higher requirements for performance or stability, including enterprise users running latency-sensitive, high-throughput AI workloads. Pricing is based on reserved computing capacity, taking into account underlying infrastructure costs, including computing resources and associated operational expenses.

*On-Premise Deployment Solutions*. We deploy our inference engine and computing resource orchestration system within the customer's own data center or private computing environment. This enables customers to use tokens across a wide range of models on their own computing resources. On-premise deployment solutions are primarily targeted at large enterprise customers and institutional clients. We generally offer such solutions through subscription-based or perpetual buyout, with fees structured as a combination of a software licensing fee (on a subscription or one-time basis), a professional implementation fee and, where applicable, a maintenance fee, each determined based on factors such as deployment scale, system configuration and customer requirements. Our on-premise deployment solutions are based on a standardized, modular platform architecture, with only limited customization required to accommodate customers' specific deployment environments or operational requirements.

Our public cloud-based services and on-premise deployment solutions form an integrated commercial system that addresses diverse customer needs. The public cloud platform enables us to serve a broad customer base, facilitate testing and validation across diverse application scenarios, and support customers' proof-of-concept and initial deployment requirements. As customers scale their usage or require greater control over performance, stability, physical location of data, network boundaries, and supply chain assurance, they may opt for dedicated instances or on-premise deployment solutions, thereby deepening engagement and expanding revenue opportunities.

## Our Commercialization and Financial Performance

Leveraging our open and independent positioning, full-stack inference optimization technologies and diversified product matrix, we have achieved rapid commercial scale-up. Following our inception in August 2023, our revenue grew significantly, surging more than sevenfold from RMB7.3 million in 2024 to RMB55.3 million in 2025.

We have established relationships with leading customers across key sectors including AI, internet, energy, finance and telecommunications, as well as ecosystem partners, such as chip designers and model developers. In the public cloud market, we have become a widely recognized AI infrastructure platform within the developer community, with token throughput ranking among the leading platforms globally. In the private cloud market, we have established deep integration with China's domestic computing ecosystem through large-scale deployments across leading domestic chip platforms, including Ascend, MetaX and Moore Threads. We became industry's first to launch DeepSeek-R1 and V3 token services based on domestic chips in February 2025, completed globally the first commercial deployment of Kimi-K2 token services on MetaX C550 chips in July 2025, and achieved state-of-the-art performance on DeepSeek-V3 using Moore Threads chips. These deployments demonstrate our ability to transform domestic AI chips from theoretical computing capabilities into commercially deployable token production capacity capable of supporting real-world AI applications at scale.

We believe our ability to combine technological innovation, large-scale engineering deployment and ecosystem collaboration positions us strongly to capture the growing demand for tokens and sustain our long-term growth trajectory.

## OUR STRENGTHS

### Open, Independent Platform Serving as a Critical Infrastructure Layer in the AI Value Chain

We are an open, independent token supply platform, providing token supply based on diverse combinations of computing resources and models with different performance-to-cost profiles. Serving a broad range of AI application developers and enterprise customers, we facilitate the scalable deployment and commercialization of AI applications and contribute to the advancement of global intelligence transformation. As of the Latest Practicable Date, we had supported a cumulative total of over 170 models, ranking among the leading independent ecosystem token supply platforms globally in terms of model coverage. Our open and independent ecosystem positioning has strengthened our global partnership network, reinforced our structural competitive advantages and established us as a critical infrastructure layer in the global AI value chain.

We believe our open positioning creates long-term strategic value within the rapidly evolving AI industry. Driven by rapid growth in AI demand, the AI value chain has attracted a broad range of participants across computing resources, models, and system software, resulting in accelerating model iteration, increasingly diversified model architectures, continuous chip innovation and multi-layered software ecosystems. As a result, interoperability and efficient coordination across chips, models and system software have become increasingly important to large-scale AI deployment. Through our open platform architecture, we effectively integrate diverse models and heterogeneous computing resources and transform them into standardized and scalable token supply services, thereby enabling broader collaboration across the AI value chain.

We are committed to maintaining an independent platform that connects the model layer and the computing layer without being tied to any single supplier. For developers, our platform provides one-stop access to diversified model capabilities and flexible deployment options. For enterprise customers, our neutral positioning enhances operational flexibility and strategic independence while enabling us to provide enterprise-grade service level guarantees. We believe our independent token supply services enable customers to achieve more flexible, efficient and reliable token services across a wide range of use cases.

Our commercialization results validate this open and independent platform strategy. The number of our customers increased from more than 2,400 in 2024 to more than 716,000 in 2025, driven by a diversified customer base spanning energy, telecommunications operators, large enterprises, and financial institutions. We believe such broad adoption demonstrates both the scalability and commercial relevance of our platform.

### Full-Stack Inference Engine With Cross-Chip, Multi-Model Adaptation Capabilities

Our proprietary inference engine is purpose-built for high-performance AI model inference. It loads pre-trained model weights and computational graphs onto underlying computing hardware, processes user prompts and efficiently generates responses through a series of system-level optimization technologies. The inference engine serves as the core execution layer that transforms computing resources into usable AI capabilities. By providing unified interfaces for model loading, execution and scheduling, our inference engine abstracts the underlying architectural differences among GPUs, NPUs and ASICs, enabling the same model service to operate efficiently across different computing platforms.

Our inference engine supports a broad range of advanced inference technologies and enables the efficient deployment of mainstream AI models, including DeepSeek, GLM, Kimi and Qwen. It is designed to maximize computing resource utilization, reduce per-token costs and deliver stable, low-latency inference performance under high-concurrency AI workloads through sophisticated request scheduling and resource management. Our inference engine also supports rapid adaptation to evolving model architectures, including dense models, sparse mixture-of-experts models and multimodal models. Supported by our streamlined model adaptation and deployment workflow covering model weight acquisition, framework adaptation, performance validation, accuracy verification and production rollout, we are able to rapidly support newly released mainstream models and maintain broad compatibility across a diverse model portfolio.

### Heterogeneous Compatibility Supporting Diverse AI Chips

We are among the few token suppliers capable of supporting both leading international chips, such as those from NVIDIA and AMD, as well as major domestic AI chips, including Ascend, MetaX, Moore Threads, Hygon and T-Head, in large-scale production environments. Leveraging our deep full-stack engineering capabilities, we transform heterogeneous computing resources into standardized, scalable tokens with flexible capacity allocation.

We have contributed to the advancement of domestic AI chips from theoretical compatibility to production-grade inference deployment, transforming domestic computing resources into commercially viable and cost-efficient token supply. For example, we collaborated with Huawei Cloud to jointly develop and optimize an inference solution for the CloudMatrix supernode cluster, achieving single-chip decode throughput exceeding 1,920 tokens/s on DeepSeek-R1. We also completed the world's first commercial service deployment of the Kimi-K2 on a cluster built on MetaX C550 chips. In addition, we achieved single-chip prefill throughput exceeding 4,000 tokens/s and decode throughput exceeding 1,000 tokens/s on the MTT S5000 chips developed by Moore Threads on DeepSeek-V3. We also continue to expand our collaboration with other domestic AI chip vendors. In particular, enabling AI model inference on domestic chips requires substantial system-level software reconstruction across the AI technology stack. We have overcome a series of complex technical challenges, including ecosystem compatibility challenges arising from tightly coupled codebases, missing operators and architectural differences; memory challenges relating to memory capacity and bandwidth bottlenecks; communication challenges caused by the lack of high-speed interconnect infrastructure and insufficient communication library optimization; and precision challenges resulting from limited low-precision computing support. Through these optimizations, we have enabled industry-leading AI model inference on domestic chips to operate reliably, efficiently and accurately at scale.

As global demand for low-cost, cross-chip and multi-model token services continues to grow, we believe the full-stack optimization and heterogeneous adaptation capabilities that we have developed in China's highly fragmented computing environment position us strongly to serve global AI demand.

**Flexible Computing Resource Orchestration and Maximized Token Production Efficiency**

We have established core competitive strengths in computing resource orchestration and token production efficiency optimization across the technology, commercial and operational dimensions of our business.

*Technology*. We have developed a production-grade computing resource orchestration system purpose-built for AI inference workloads. Building upon the virtualization, scheduling and delivery capabilities of traditional cloud infrastructure, we have further optimized our infrastructure stack for the unique characteristics of AI inference workloads, including heterogeneous computing resources, large-scale models, low-latency inference and burst concurrency demands. We have systematically addressed a broad range of engineering challenges relating to computing, networking, storage, scheduling, security and observability. Specifically, our secure container architecture replaces the conventional shared-kernel container model with a virtual machine-level runtime environment, enabling near hardware-level isolation and enhancing tenant data privacy and operational security. Our distributed model distribution and caching system pools storage and bandwidth resources across cluster nodes, enabling efficient storage and rapid distribution of large-scale model weights. In addition, our AI model API gateway aggregates multiple inference clusters into a unified service endpoint capable of dynamically routing requests across heterogeneous computing resources, monitoring backend system conditions in real time and performing automatic failover, thereby ensuring high service continuity.

Built on top of this infrastructure foundation, our proprietary inference engine reduces inference latency by up to 70% and improves throughput by three to five times, while our dynamic quantization technology reduces inference computation requirements by 60% to 80%, maximizing token production efficiency per chip. Our overall infrastructure architecture delivers a 99.95% SLA and supports automatic failover, circuit breaker and request-throttling capabilities, enabling stable operation of production-grade, high-concurrency AI applications.

*Commercial*. We have developed diversified business models.

- For developers and enterprise customers, we have supported a cumulative total of over 170 models and industry standard protocols, enabling us to rapidly attract customers through competitive pricing and broad ecosystem coverage.

- For enterprises and institutions with self-built computing clusters, we provide solutions that improve inference efficiency while enabling customers to monetize surplus computing capacity through external token services.

- In addition, we adopted a joint operations model, under which we collaborate with regional intelligent computing centers by providing our technology platform and brand support, while partners participate in revenue sharing based on actual token service volume, significantly lowering their technology investment and operational barriers.

Growing customer usage attracts additional model providers and computing resource partners to our platform, while broader model coverage and computing resource availability further enhance customer adoption and platform activity.

*Operational*. We have established operational capabilities spanning the full lifecycle from computing resources to scalable token supply. On the demand side, leveraging our broad model ecosystem and industry-focused solutions, we serve customers across sectors including AI, internet, energy, finance, and telecommunications, helping improve computing resource utilization and generate stable demand for our computing resource partners. On the supply side, through visualized management tools and more than 30 ready-to-use deployment templates, we reduce computing resource integration time to under three minutes without requiring manual configuration, significantly lowering onboarding complexity for computing resource providers. By connecting fragmented computing resources with diversified token demand through our platform, we continuously improve resource utilization efficiency and strengthen coordination across the AI ecosystem.

**Strong Customer Acquisition and Retention Driven by Product Capabilities and Diverse Application Scenarios**

As an early participant in the token supply market, we established broad influence within the AI community before global demand for token consumption accelerated at scale. Leveraging our first-mover advantages and broad model adaptation capabilities, we have attracted and retained a large base of sophisticated developers, emerging AI-native companies and enterprise customers. Through continuous platform usage and integration, we have developed strong customer stickiness and accumulated significant market recognition.

Our platform is designed to maximize the value of different combinations of chips and models, enabling customers to select the most cost-efficient token service for specific application scenarios based on their requirements for performance, latency, quality and pricing. By orchestrating heterogeneous computing resources and supporting a broad range of models, we are able to provide customers with optimized token supply across diverse AI application scenarios, including coding, AI agent, multimodal generation and enterprise AI applications. Different workloads require different trade-offs between token generation speed, model capability and cost efficiency, and our platform enables customers to dynamically select the most suitable solution for their production needs.

We provide customers with comprehensive solutions spanning the full deployment spectrum, including serverless token services, dedicated instances and on-premise deployment solutions designed for customers with stringent compliance, security and performance requirements. Our public cloud services drive developer ecosystem growth and platform-scale effects by providing convenient access to a broad range of models without requiring customers to manage complex AI infrastructure. Customers may choose among serverless, dedicated or on-premise deployment service modes based on their operational requirements and cost considerations, while also benefiting from flexible deployment options and low-latency token services. At the same time, our on-premise deployment solutions provide enterprise customers with greater data isolation, operational control and customized performance optimization capabilities, contributing higher-margin revenue streams and stronger enterprise customer retention. This diversified product portfolio enhances customer conversion across different stages of AI adoption and allows us to maintain strong and resilient customer retention as customers scale from proof-of-concept projects to large-scale production deployments.

In addition, feedback and operational insights generated from our high-quality customer base continuously improve our understanding of customer needs and application scenarios, enabling us to optimize our platform capabilities, model selection and resource orchestration strategies. We believe this creates a strong flywheel effect that enhances customer stickiness, service quality and ecosystem competitiveness over time.

**A Strong Engineering Culture Rooted in Technical Excellence**

Our deep technical expertise and engineering-driven culture form the foundation of our competitive strengths. Our founder, Dr. Yuan Jinhui, is a seasoned AI entrepreneur with multiple successful ventures in the AI sector. From our inception, we identified token supply as a critical opportunity in the AI value chain based on our conviction that scalable, efficient and high performance-to-cost inference would become a key bottleneck as AI models unlock diverse application scenarios and token demand continues to grow rapidly.

We have assembled a highly specialized engineering team with expertise spanning distributed systems, inference engines and cloud-native infrastructure. Our core capability lies in efficiently adapting rapidly evolving model architectures across heterogeneous computing resources and optimizing system-level performance across latency, throughput and cost efficiency. This end-to-end engineering capability enables us to complete inference engine adaptation and performance optimization for domestic chips within approximately three months, significantly faster than traditional cloud vendors according to Frost & Sullivan. Our proprietary inference engine, dynamic quantization algorithms and distributed caching systems were developed through engineer-led innovation and rapidly deployed into production

environments, demonstrating our strong engineering execution capabilities. We maintain a flat and engineering-centric organizational structure in which major technical decisions are driven by teams with frontline engineering experience and validated through rigorous technical evaluation and data-driven analysis.

Our management team combines deep engineering expertise from leading global cloud computing and AI companies with strong strategic execution capabilities. We identified production-grade inference infrastructure as a strategic focus at an early stage. Through continuous technological investment and long-term engineering accumulation, we have established differentiated capabilities and deep engineering know-how across heterogeneous computing adaptation, inference optimization and large-scale resource orchestration. We believe the cumulative and compounding nature of these capabilities has created durable long-term competitive advantages in the AI model inference infrastructure market.

## OUR STRATEGIES

To strengthen our competitive positioning and support our long-term growth in the AI era, we intend to pursue the following key strategies:

### Enhance Token Production Efficiency and Broaden Token Application Use Cases Through Sustained Investment in Research and Development

As AI models continue to evolve rapidly and AI infrastructure continues to upgrade, we believe customers' growing demand for high-quality tokens will remain a key driver of our rapid business growth. We plan to continue investing in the research and development of our core technology foundation to further enhance efficiency across a broad range of heterogeneous chips. At the same time, we seek to deepen collaboration across the AI ecosystem and continuously expand token application scenarios for developers and enterprise customers globally. In particular, we intend to:

- *Enhance our model adaptation capabilities*. We plan to continue investing in our ability to adapt to and integrate into newly released AI models and further broaden the range of models supported by our platform. We also maintain close communication and business cooperation with model providers in order to remain aligned with emerging AI technology trends and developments;

- *Strengthen compatibility with heterogeneous chips*. We plan to continue expanding and optimizing support for a wide range of domestic and international chips and transform heterogeneous computing resources into high-performance, standardized token services with flexible resource scheduling capabilities. We also intend to further strengthen optimization capabilities for domestic chips in response to the constrained supply environment of AI computing resources;

- *Continue to improve our computing resource orchestration system*. We plan to continue investing in the research and development of our computing resource orchestration system to further enhance its scalability, reliability and security while strengthening its capabilities across computing, storage, networking, scheduling and infrastructure management. As AI inference workloads continue to evolve, we seek to further optimize our cloud software for heterogeneous computing environments and increasingly complex enterprise AI deployment scenarios;

- *Advance next-generation inference technologies*. We plan to advance a suite of next-generation inference technologies, including (i) a decentralized, dataflow-based inference architecture that replaces conventional layer-centric coordination with lightweight, on-chip processing units operating over high-speed interconnects for low-latency, high-throughput and fault-tolerant inference, and (ii) a large-scale context memory system that aggregates cluster-wide memory and storage into a scalable, fault-tolerant memory pool, enabling efficient storage and retrieval of near-infinite context for next-generation AI applications;

- *Improve token production efficiency across both the application and infrastructure layers*. At the application layer, as AI agents become a major driver of token consumption and enterprise intelligence transformation, we plan to build efficient AI agent middleware frameworks to support the development of innovative AI-native applications with lower token consumption; and

- *Continue building an open token ecosystem*. We will strengthen our position as an open and independent platform with full-stack technological capabilities and seek to promote broader adoption of our model inference and resource orchestration technologies across the AI ecosystem. In the public cloud domain, we seek to more efficiently orchestrate heterogeneous computing resources and continuously expand the scale of computing resources accessible through our platform. In the private cloud domain, we aim to provide one-stop solutions that help more enterprises unlock the value of their internal data and computing infrastructure and establish enterprise-level "token factories."

**Expand Our Coverage of Key Customers and Enhance Our Market Presence**

We intend to continue expanding our customer coverage, particularly among customers with significant token consumption potential. Leveraging our core technological capabilities, we also seek to deepen our penetration among customers that (i) require flexible access to multiple AI models, (ii) have substantial demand for domestic AI computing infrastructure and place significant emphasis on AI and data security, and (iii) possess heterogeneous computing infrastructure and seek to improve computing resource orchestration efficiency and utilization of existing computing assets.

We plan to continue strengthening our brand recognition and market presence. In particular, we intend to continuously enhance our product functionalities, expand our product offerings and strengthen our service capabilities in order to further improve market recognition and brand influence. We also intend to publish technical white papers, case studies, benchmarking reports and industry solutions to demonstrate our technological capabilities and service strengths and further enhance our professional reputation and market credibility. In addition, we plan to participate in AI infrastructure industry conferences, forums and technical events through keynote presentations, product demonstrations, joint solution launches and customer case sharing, thereby increasing brand visibility and improving customer acquisition and conversion efficiency.

We further intend to continue enhancing our integrated online and offline sales and marketing capabilities. In particular, we plan to expand our channel sales model through customer acquisition initiatives via our official website and developer platforms, online marketing activities, technical community development and content marketing. We also intend to strengthen our direct sales capabilities targeting key industries and large enterprise customers by expanding our sales and solution teams with experience in AI infrastructure, cloud computing, enterprise software and industry solutions.

**Strengthen Computing Resource Supply Capabilities and Supply Chain Resilience**

As AI applications continue to expand, we believe the supply of AI computing resources is likely to remain constrained in the near to medium term. We therefore intend to continue strengthening our computing resource supply capabilities and supply chain management in order to enhance the stability of our business operations and mitigate the impact of fluctuations in computing resource pricing. In particular, we plan to improve the stability and flexibility of our access to computing resources, expanding compatibility with a broader range of AI computing chip models and specifications, and continuously strengthening adaptation capabilities for domestic AI chips. Through these initiatives, we seek to further diversify and expand our computing resource supply and enhance the resilience of our supply chain.

We also intend to further deepen our strategic cooperation with key computing resource suppliers and explore additional cooperation models, including joint computing operations and computing resource management arrangements, in order to improve resource acquisition efficiency, utilization efficiency and overall operational coordination. We believe these initiatives will further strengthen our computing resource supply capabilities, support our continued business growth and long-term development, and enhance our position within the broader AI infrastructure ecosystem.

**Advance Our Global Expansion Strategy by Leveraging Overseas Ecosystems and Our Technological Advantages**

We intend to continue expanding into overseas markets and leverage the full-stack technological capabilities and product strengths we have developed in China to enhance our competitiveness globally. We believe overseas markets present attractive opportunities for token services due to abundant advanced computing resources, mature AI application ecosystems and increasing international adoption of domestic AI models driven by their performance-to-cost advantages. We plan to advance our international expansion strategy in a phased manner, with product capabilities serving as the primary driver of overseas growth. In particular, we intend to:

- Expand our overseas customer base through our token production efficiency, broad model support capabilities and integrated product offerings, together with overseas online platforms and digital marketing channels.

- Strengthen localized ecosystem partnerships by collaborating with local ecosystem partners to better serve customers requiring localized support.

- Progressively establish localized sales and operational teams in selected overseas markets in order to enhance customer service capabilities, market responsiveness and local delivery capabilities.

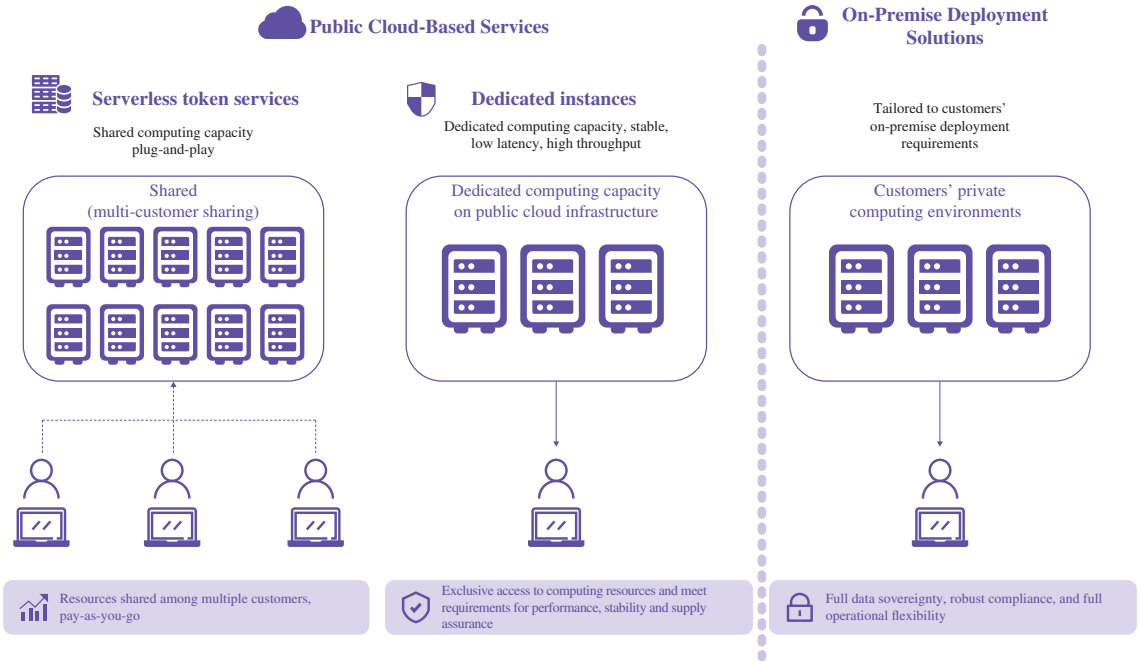
**Cultivate a Globally Competitive Talent and Management Platform**

We intend to continue strengthening our talent and management capabilities in key areas such as model inference and AI-native cloud infrastructure. In particular, we plan to attract experienced industry professionals, interdisciplinary talent and internationally oriented commercialization personnel across technology, management, marketing and other functions in order to support our continued technological innovation and business expansion. We highly value employees' technological expertise, industry insights, practical experience and ability to adapt to emerging AI technologies, and seek to build a high-caliber team capable of supporting our long-term development in the AI-native era. We also intend to continue fostering a fair, open and innovation-oriented organizational environment and corporate culture that provides employees with opportunities for growth and development. We believe these efforts will help us attract, develop and retain talent aligned with our business needs and support our continuing technology innovation, product optimization and business growth.

## OUR PRODUCTS AND SERVICES

Our business primarily comprises two complementary product and service lines designed to address a wide range of customer deployment and usage needs: public cloud-based services and on-premise deployment solutions. Our public cloud-based services include (i) serverless token services, through which customers access and utilize a wide selection of AI models on a usage basis, and (ii) dedicated instances, under which we offer reserved computing capacity on public cloud infrastructure for customers requiring enhanced performance, stability, and customization. Under on-premise deployment solutions, we primarily provide system software that enables customers to deploy and operate our platform within their own infrastructure environments.

Our public cloud-based services and on-premise deployment solutions form an integrated commercial system that addresses diverse and evolving customer needs. The public cloud platform enables us to serve a broad customer base, support proof-of-concept and initial deployment use cases, and facilitate validation across diverse application scenarios. As customers grow their usage or require greater control over performance, stability, physical location of data, network boundaries, and supply chain assurance, they may opt for dedicated instances or on-premise solutions — deepening engagement and expanding revenue opportunities across the customer lifecycle.




The following table sets forth our revenue by business line for the periods indicated:

| | For the Period from August 29 to December 31, | | For the Year Ended December 31, | | | |
|---|---|---|---|---|---|---|
| | 2023 | | 2024 | | 2025 | |
| | RMB | % | RMB | % | RMB | % |
| | *(in thousands, except for percentages)* | | | | | |
| Public cloud-based services | – | – | 1,069 | 14.6 | 29,261 | 52.9 |
| On-premise deployment solutions | 6 | 100.0 | 6,277 | 85.4 | 26,069 | 47.1 |
| **Total** | **6** | **100.0** | **7,346** | **100.0** | **55,330** | **100.0** |

## Public Cloud-Based Services

Our public cloud-based services are designed to deliver standardized, high-quality tokens at scale through a single standard interface supporting a cumulative total of over 170 models. The services support a broad range of application scenarios — including coding, agents and multimodal applications — providing customers with unified, flexible access to AI capabilities. We launched our serverless token services in May 2024, marking the commencement of large-scale commercialization. The platform supports two complementary service offerings: serverless token services on a shared resource basis and dedicated instances providing reserved computing capacity, which together address the full spectrum of customer requirements, from individual developers and startups seeking cost-efficient, on-demand access, to large enterprises requiring dedicated performance, consistent stability or supply assurance.

***Serverless Token Services***

We provide token services on a shared-resource basis, charging fees according to token consumption. Under this model, multiple customers simultaneously utilize the same shared pool of underlying computing resources.

Customers may access our serverless token services through either (i) direct API integration, where they integrate directly with our API to access models on our platform; or (ii) Bring Your Own Key (BYOK), where customers generate an API key on our platform and use it within compatible third-party tools and applications, allowing them to retain their existing tools and workflows while selecting our platform as their token provider. As of the Latest Practicable Date, users can access models through most independent applications and development tools, including Claude Code, Hermes Agent and OpenClaw; in addition, nearly 100 of the widely used applications have pre-selected our platform as their built-in model service provider, such as LangChain, TRAE, Dify and Cherry Studio, enabling users to access our services without additional configuration.

Key features of our serverless token services are:

- *Flexible, commitment-free access.* Customers can access and switch between a broad ecosystem of leading models through a single unified interface, with no upfront infrastructure investment or long-term commitment required. Usage scales on demand, enabling customers to adapt model selection to evolving application needs.

- *Streamlined onboarding and rapid time-to-deployment.* Our standardized onboarding process – account registration, credit purchase, API key generation and application integration – is designed to minimize deployment time and enable full operational readiness within minutes.

- *Adaptable to diverse application scenarios.* Our platform supports a wide range of use cases and integration patterns, enabling customers to embed AI capabilities across varied workflows, environments and business contexts without rebuilding infrastructure for each deployment.

- *Cost-efficient, token-based pricing.* Through optimized orchestration and heterogeneous computing support, our platform improves utilization rates and reduces token costs, enabling competitive pay-as-you-go token pricing with no idle-capacity overhead.

- *End-to-end model lifecycle support.* Our platform, with its inference capabilities, supports model selection, integration and deployment, allowing customers to manage the full model customization and scaling lifecycle within a single environment – reducing reliance on disparate third-party tools.

- *Data protection by design.* Our platform is built with enterprise-grade data protection as a baseline, including non-retention of customer data during inference and flexible deployment options that give customers visibility and control over how their data is handled.

- *Bring Your Own Key (BYOK) compatibility*. Our platform supports a BYOK model under which end users of third-party AI tools and applications can configure our API key directly within those tools to access models on our platform, without modifying their existing workflows or switching applications. This allows users to freely select and switch between models available on our platform from within the tools they already use.

Our serverless token services are primarily targeted at customers seeking flexible and cost-effective access to AI capabilities, particularly developers and small- and medium-sized enterprises. Our customer base for serverless token services is highly diversified, encompassing internet companies, model developers, enterprises across a broad range of sectors and individual developers.

We generally adopt either (i) a prepaid model for customers, under which customers purchase credits in advance and usage is deducted based on tokens consumed, or (ii) a postpaid arrangement for certain recurring enterprise customers, under which we offer monthly settlement. Our systems generate itemized billing records specifying the model accessed, the volume of tokens consumed, and the corresponding charges. Input and output tokens are priced separately, and pricing may further vary depending on the model accessed and the length of the context. Pricing is generally determined with reference to prevailing market practices for token services and the pricing strategies of model providers, and we may offer pricing at a discount to such benchmark pricing, taking into account our cost structure and operational efficiencies. For large-volume customers, we may offer negotiated pricing adjustments. The shared-resource pool architecture is well-suited to token-based billing, as the underlying infrastructure costs can be spread across multiple customers, enabling competitive, usage-proportional pricing.

***Dedicated Instances***

We offer reserved computing capacity on public cloud infrastructure to customers, providing them with exclusive or prioritized specified resources.

Key features of our dedicated instances services are:

- *Exclusive computing resources with no resource sharing*. Dedicated computing resources are reserved for customers and are not shared with any other customers, ensuring complete separation of data and workloads. This allows predictable throughput and stable performance free from resource contention.

- *Template-based self-service configuration*. Customers can configure computing capacity, model selection and deployment architecture through a library of curated best-practice templates, enabling flexible self-service customization without requiring deep infrastructure expertise. Additional templates can be made available to accommodate specific customer requirements.

- *Enterprise SLA and service reliability*. Reserved computing resources enable enhanced availability, system reliability and performance consistency, supporting the requirements of mission-critical enterprise applications.

- *Seamless upgrade path from serverless token services*. Customers may begin with serverless token services for testing and early-stage deployment and transition to dedicated instances as usage scales or performance requirements increase, with no disruption to existing integrations.

Dedicated instances are typically offered to customers with higher requirements for performance, stability or supply assurance, as well as those with large and steady demand for computing resources, including enterprise users running latency-sensitive, high-throughput AI workloads. As of the Latest Practicable Date, our dedicated instances customers included leading technology and internet companies, global pharmaceutical enterprises and top-tier AIGC companies.

Dedicated instance customers may choose either (i) a fixed subscription fee for the reserved period, or (ii) usage-based charges measured in tokens consumed within the exclusive instance, depending on their preference. For prospective dedicated instance customers, we work closely with their technical and procurement teams to analyze their usage patterns–including average daily token volumes, peak throughput requirements, and the variability of their usage profile–to assist them in determining the most cost-effective configuration. Depending on the analysis, we may recommend a fixed dedicated instance, a flexible orchestration with elastic expansion capabilities, or a hybrid model. As a customer's usage grows and their requirements for greater performance and service stability increase, they may transition from our serverless token services to dedicated instances.

### On-premise Deployment Solutions

We also provide on-premise deployment solutions, where we deploy our inference engine and computing resource orchestration system within the customer's own data center or private computing environment. Our platform, once deployed within a customer's environment, effectively creates a private-cloud-like AI infrastructure within the customer's data center, enabling staff to access a diverse range of AI models through a centralized, managed interface while ensuring that all data and inference activity remains entirely within the customer's own infrastructure. Our on-premise deployment solutions are built on a standardized, modular platform architecture, with only limited customization required to accommodate customers' specific deployment environments or operational requirements.

Key features of our on-premise deployment solutions are:

- *Customer-owned infrastructure and full data sovereignty*. Computing resources are provided and operated entirely within the customer's own infrastructure. Customers retain complete control over their data, workloads and hardware, making this solution particularly suited to regulated industries and use cases involving sensitive or confidential information. At the same time, customers benefit from the full capabilities of our token supply platform without the need to independently build or maintain the underlying software stack, combining the data control of a private deployment with the convenience of a managed service.

- *Heterogeneous computing support, with domestic chip compatibility*. Our platform supports a broad range of hardware architectures, with particular emphasis on domestic chip ecosystems. This enables customers to leverage their existing or planned heterogeneous computing environments – including domestic chips – while achieving high-performance inference.

- *User-friendly console for non-technical operators*. Our platform provides visualization and self-service operational tooling designed to enable non-technical personnel to monitor, manage and operate model deployments, reducing dependency on specialized infrastructure expertise and accelerating adoption across the organization.

Our on-premise deployment solutions are typically adopted by large enterprise customers and institutional clients with heightened requirements for physical location of data, network boundaries or supply assurance, or system integration, and may be delivered in the form of software licenses and/or integrated software-hardware solutions depending on the scale and complexity of deployment. As of the Latest Practicable Date, our on-premise deployment solutions customers included major players in the AI, internet, energy, finance and telecommunications industries.

We generally offer such solutions through subscription-based or perpetual buyout sales models, with fees typically structured as a combination of a software licensing fee, a professional implementation fee, and, where applicable, a maintenance fee, each determined based on factors such as deployment scale, system configuration and customer requirements. Subscription arrangements are available for customers who prefer to align software expenditure with annual budgeting cycles or who wish to reassess their platform requirements at regular intervals given the rapid pace of AI model development. Following delivery, we provide ongoing maintenance and support services. On-premise deployments are typically completed within a couple of weeks.

## Key Operating Metrics

The following table sets forth our key operating metrics for the year/period indicated.

| | As of/For the Year Ended December 31, | | As of/For the Four Months Ended April 30, |
|---|---|---|---|
| | 2024 | 2025 | 2026 |
| **Public cloud-based services** | | | |
| – Average daily token throughput in last month of the period (in billions) [(1)] | 47.8 | 163.1 | 578.5 |
| – Number of registered users [(2)] | 127,133 | 9,197,439 | 10,282,431 |
| ***– Serverless token services:*** | | | |
| – Number of active users[(3)(4)] | 57,431 | 5,452,952 | 1,453,923 |
| – Number of serverless token services customers[(3)(5)] | 2,455 | 716,012 | 642,866 |
| ***– Dedicated instances:*** | | | |
| – Number of dedicated instance customers[(3)] | 7 | 49 | 20 |
| – Average revenue per dedicated instance customer (in RMB) | 87,669 | 305,353 | N/A |
| **On-premise deployment solutions** | | | |
| – Number of on-premise deployment customers[(3)] | 28 | 20 | 5 |
| – Average revenue per on-premise deployment customer (in RMB) | 224,184 | 1,303,448 | N/A |

*Notes:*

* The operating metrics as of/for the year ended December 31, 2023 was not available as our Company was incorporated in August 2023 and did not generate material revenue in 2023.

(1) Average daily token throughput in the last month of the period is calculated by dividing the total number of tokens processed in that month by the number of calendar days in the month. For this purpose, "tokens" include both input tokens submitted by users and output tokens used by the models, both produced by us.

(2) Represents a cumulative number as of the period end.

(3) Represents the numbers in 2024 and 2025 and the four months ended April 30, 2026. The number for the four months ended April 30, 2026 was not annualized and therefore was lower.

(4) Active users refer to the users who used our serverless token services in the year/period indicated.

(5) Serverless token services customers refer to the users who purchased our serverless token services in the year/period indicated.

Our key operating metrics reflect the rapid expansion of our business across both our public cloud-based services and on-premise deployment solutions during the Track Record Period and into 2026.

***Public Cloud-Based Services.*** Our platform experienced rapid growth in both user scale and token throughput. Our average daily token throughput grew by 241.1% from December 2024 to December 2025, and by a further 254.7% from December 2025 to April 2026. Our cumulative registered user base expanded by approximately 72-fold from December 31, 2024 to December 31, 2025, and continued to grow into 2026, reflecting broadening adoption of our platform driven by the rapid proliferation of AI applications and the expanding ecosystem of developers and enterprises utilizing token supply services. Specifically:

- For our serverless token services, the number of active users increased by approximately 95-fold from 2024 to 2025, while the number of paying customers grew by approximately 292-fold over the same period. This growth was primarily driven by the increasing adoption of AI applications across industries and the growing recognition of our platform's capabilities,

which together attracted a significant influx of developers and enterprise users. The number of paying customers for the four months ended April 30, 2026 approached the full-year 2025 level, underscoring the continued strong momentum in user monetization into 2026.

- For our dedicated instances, the number of customers increased by 600.0% from 2024 to 2025, as enterprise customers with growing token consumption and heightened requirements for performance stability and data isolation transitioned into reserved computing resources. Average revenue per dedicated instance customer increased by 248.3% from 2024 to 2025, reflecting deeper commercial engagement and higher contract values as our enterprise relationships matured.

***On-Premise Deployment Solutions.*** Our on-premise deployment solution is project-based in nature. The number of on-premise deployment customers decreased by 28.6% from 2024 to 2025, while average revenue per customer increased significantly by 481.4% over the same period, reflecting a shift toward larger-scale enterprise deployments with higher contract values.

**Case Study**

***Case Study 1: Serverless token services — AI Smart Toy Company***

*Customer Background:* The customer is an innovative company specialising in AI-powered children's toys, focused on building intelligent, interactive companion products. Its toys rely on real-time understanding of children's voice commands and conversational intent.

*Customer Needs:* Children are highly sensitive to response delays — even a few seconds of lag is enough to disengage them. The customer therefore needed inference that could deliver millisecond-level responses consistently. Children's speech also presents particular challenges, including unclear pronunciation, non-standard grammar and emotionally driven expression, placing high demands on the model's intent recognition capability. Beyond performance, the customer needed a cost structure that could flex with demand: usage spikes during school holidays and quieter periods overnight made fixed infrastructure investment inefficient.

*Our Solution:* The customer selected our serverless token services as the AI foundation for its products. We provided the customer with access to all available models on our platform via a single API call, with new model releases typically available within a week, allowing the customer to stay current with frontier capabilities without any independent deployment work.

*Value Delivered:* After adopting our serverless token services, the customer saw measurable gains. Average product response time fell from seconds to milliseconds; under high-concurrency conditions, latency dropped below 500ms, directly improving children's interactive experience and driving a doubling of demand volume. The pay-as-you-go model eliminated idle resource spend and gave the customer the flexibility to scale effortlessly through peak periods such as school holidays without over-provisioning. With new models accessible instantly, the team was able to test and validate model capabilities rapidly and translate insights into product improvements without carrying the burden of infrastructure operations, freeing engineering effort for product innovation and user experience optimization.

***Case Study 2: Dedicated Instance — AI Coding Platform***

*Customer Background:* The customer is an AI company in China specializing in full-stack coding workflows, covering code design, writing, completion and testing, all of which rely on real-time AI model inference. The customer had been using our serverless token services since the early stages of its product development. As its AI agent products expanded in deployment, its daily token consumption reached the hundred-billion level to maintain prompt access to the latest model.

*Customer Needs:* As consumption scaled, the customer's existing serverless token arrangement became less suited to its production requirements. During peak periods, computing resources were subject to contention, creating reliability risks for production workloads. At the same time, the per-token pricing structure became less cost-effective as usage grew larger and more predictable. The customer needed a way to secure stable, dedicated computing capacity with guaranteed performance and a cost structure better suited to its scale without taking on the overhead of managing infrastructure.

*Our Solution:* We transitioned the customer's core inference workloads to our dedicated instance service, providing exclusive computing resources for its production operations. Since the dedicated instance service uses the same API interface as our serverless token services, the migration required no application-level changes. The reserved computing resources are not shared with any other customer, which resolved the peak-period contention issue, and the service is backed by enterprise-grade SLA commitments.

*Value Delivered:* Following the transition, the customer saw improvements in stability, performance and cost. With resource contention eliminated, the customer achieved 99.95% SLA availability, enabling its hundred-billion-level daily token workloads to run without disruption. The customer also achieved over 50% cost reduction through tailored instance configurations combined with a cache hit rate exceeding 90%. The customer has since continued to scale its reserved capacity in line with growing business volumes, with its engineering teams freed to focus on product development rather than infrastructure operations.

***Case Study 3: On-premise Deployment Solutions — State-Owned Power Enterprise***

*Customer Background:* The customer is one of China's largest state-owned power enterprises, operating across electricity generation, transmission and renewable energy. As AI adoption accelerated across the industry, the customer sought to establish a comprehensive AI capability covering a wide range of use cases, from intelligent office operations and industrial safety management to renewable energy operations and maintenance.

*Customer Needs:* The customer had no unified AI platform, resulting in fragmented processes, low resource utilization and rising costs. Its production use cases demanded high throughput and low latency that open-source solutions alone could not meet. It also needed a solution that could accommodate its heterogeneous computing resources without significant custom adaptation work, and integrate the latest frontier models quickly without being held back by traditional deployment cycles.

*Our Solution:* We deployed our token supply platform on-premise, giving the customer a unified environment for model training, inference, knowledge base management and agent orchestration.

*Value Delivered:* Following deployment, the customer's heterogeneous computing resources — including domestic chips — were consolidated under a single managed platform, converting idle capacity into productive use and materially simplifying infrastructure. The platform had access to nearly 100 models, which are continuously updated, while inference performance improved substantially with DeepSeek V3 alone saw a boost of over 50%. Across operations, the platform enabled intelligent office workflows, enterprise knowledge base construction, and AI agent development, reducing administrative burden and improving day-to-day efficiency. Notably, intelligent wind turbine diagnostics now complete in a third of the time, with accuracy rising from approximately 75% to over 95%.

## ELIGIBILITY ASSESSMENT UNDER CHAPTER 18C FRAMEWORK

We are seeking [**REDACTED**] under Chapter 18C of the Listing Rules. We are primarily engaged in the design, development and commercialization of a token supply platform. All of our products and services are designated Specialist Technology Products as defined under Chapter 18C of the Listing Rules. Our Directors are of the view that our products and services fall within an acceptable sector of a Specialist Technology Industry as defined under Chapter 18C of the Listing Rules on the following basis, as advised by Frost & Sullivan.

The table below sets out a summary for how our products and services fall within the Artificial intelligence ("AI") sector, an acceptable sector under the Specialist Technology Industries defined in Chapter 18C of the Listing Rules.

| Specialist Technology Products | Specialist Technology Industry Acceptable Section | Main Function Analysis | Business Model | Timeline of Commercialization and Stage of R&D |
|---|---|---|---|---|
| Public cloud-based services | Artificial intelligence ("AI") (technology and infrastructure enabling AI) | Public cloud-based services are designed to deliver standardized, high-quality tokens at scale, providing customers with unified and flexible access to a cumulative total of over 170 models through a single, standard interface, covering a broad range of AI application scenarios including coding, agents and multimodal applications. | Combination of consumption-based and subscription based | Commercialization since May 2024 with ongoing R&D |
| On-premise deployment solutions | Artificial intelligence ("AI") (technology and infrastructure enabling AI) | On-premise deployment solutions are designed to 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. | Combination of transaction-based and subscription-based | Commercialization since December 2023 with ongoing R&D |

For further details, see "— Our Products and Services."

## PATH TO COMMERCIALIZATION

### Commercialization Status

The following table illustrates the key commercialization timeline of our Specialist Technology Products.

| Date | Milestone |
|---|---|
| December 2023 | We started to generate revenue through software subscriptions, marking the beginning of our commercial operations. According to Frost & Sullivan, our commercialization timeline outpaced the industry average. |
| May 2024 | We launched our serverless token services, marking the transition to large-scale commercialization of our public cloud-based services. |
| July 2024 | We launched our dedicated instances services, providing enterprise customers with reserved computing resources for production-grade AI workloads. |
| February 2025 | We became the first to deploy DeepSeek-R1 & V3 on domestic chips. This achievement represented the first successful adaptation of domestic chips for token services in the industry, demonstrating the commercial viability and usability of domestic chip architectures for production-grade AI workloads and significantly elevated our public profile. We commenced large-scale commercial deployment of cross-chip and multi-model adaptation. |
| April 2025 | We started offering our full-stack on-premise deployment solutions to enterprise customers. |
| June 2025 | We officially launched our serverless token services for international customers, making our services accessible to overseas developers and enterprises. |
| August 2025 | Our serverless token services were listed on a leading AI model routing platform. This represented a further expansion of our international business. |
| December 2025 | The number of registered users on our platform exceeded nine million and the number of our enterprise customers exceeded 10,000. |
| April 2026 | Our peak daily token throughput exceeded one trillion for the first time, and the number of registered users on our platform exceeded 10 million, demonstrating continued growth in developer and enterprise adoption of our token services. |
| May 2026 | Our overseas monthly revenue exceeded USD1 million. |

We expect that we will satisfy the revenue requirement for a commercial company defined under Chapter 18C of the Listing Rules ("**Commercial Company**") by the end of 2026.

### Commercialization Assumptions

Our projection of becoming a Commercial Company by the end of 2026 is subject to significant uncertainties, including but not limited to the following: (i) there will be no material delays or obstacles affecting our commercialization plans, product and service development roadmap or go-to-market execution; (ii) we will be able to continue to develop, deploy and deliver our services and solutions in a timely manner and with the quality and performance expected by customers; (iii) our customers and prospective customers will continue to accept and adopt our products and services, and complete their internal evaluation, testing, budget approval and procurement processes, at a pace broadly consistent with our expectations; (iv) we will be able to maintain sufficient access to computing resources from our suppliers and other infrastructure partners, as well as sufficient technical, operational, supply chain, compliance and quality control capabilities, to support our commercialization growth; (v) there will be no material adverse changes in the regulatory or policy environment affecting our services, computing resource procurement, overseas operations or commercialization activities; (vi) market demand for token services will continue to grow and will not decline materially due to macroeconomic, technological or industry-specific conditions; (vii) we will be able to continue to attract and retain key personnel in R&D, infrastructure operations and commercialization functions; and (viii) there will be no occurrence of any material adverse event described in "Risk Factors." Accordingly, while we believe we have a reasonable basis for our expectation regarding the timeline to satisfy the revenue requirement for a Commercial Company, if any of the foregoing assumptions proves to be inaccurate, our commercialization may be delayed and our revenue growth may be adversely affected. Save for the assumptions set out above, we did not foresee any material impediment to achieving our revenue target as of the Latest Practicable Date.

## BUSINESS SUSTAINABILITY

### Historical Loss-Making

We are a Pre-Commercial Company defined under Chapter 18C of the Listing Rules. During the Track Record Period, we successfully transitioned to the initial commercialization stage, experiencing rapid top-line growth. Our revenue increased significantly from RMB6 thousand in 2023 to RMB7.3 million in 2024, and further increased by 653.2% to RMB55.3 million in 2025, primarily driven by the expansion of our customer base across both our public cloud-based services and on-premise deployment solutions. Despite this rapid revenue growth, during the Track Record Period, we recorded a loss for the period/year of RMB12.2 million, RMB81.9 million and RMB345.5 million in 2023, 2024 and 2025, respectively, alongside an adjusted net loss (non-IFRS measure) of RMB12.2 million, RMB54.0 million and RMB187.1 million for the same periods, primarily due to the following reasons:

***Gross margin pressure from early-stage commercialization.*** We are an emerging company in the token supply industry, and the commercialization of our platform and services remains at an early stage. Our cost of revenue, which primarily comprises computing resource costs in the form of computing power rental fees, grew at a significantly faster rate than our revenue during our early commercialization phase, increasing from RMB1 thousand in 2023 to RMB4.5 million in 2024, and further by 1,441.6% to RMB68.6 million in 2025, with computing resource costs representing 86.9% of total cost of revenue in 2025. Consequently, our overall gross margin shifted from 83.3% in 2023 to 39.4% in 2024, and to (24.0)% in 2025, resulting in a gross loss of RMB13.3 million in 2025. This margin contraction was primarily driven by substantial costs incurred from leasing computing resources as we continued to improve our products and services, while utilization of such computing resources was gradually ramping up. More specifically, this margin contraction and the growth of our cost of revenue outpacing our revenue were attributable to the following factors: (i) a structural shift in our revenue mix, as our public cloud-based services, which operate at a lower gross margin, accounted for a significantly larger percentage of our total revenue in 2025; (ii) negative gross margin due to the substantial computing resource costs, particularly computing power rental fees for public cloud-based services, required to secure sufficient capacity and maintain high service availability for our rapidly expanding user base, which outpaced our initial revenue generation during the early commercialization phase, and (iii) the decrease of gross margin for on-premise deployment solutions, as we onboarded new enterprise customers across a broader range of industries, which entailed a higher degree of deployment complexity during the initial onboarding phase. As a result, our cost levels were relatively high during the Track Record Period.

***Substantial investment in R&D.*** During the Track Record Period, we made substantial investments in R&D to support the development of our Specialist Technology Products, with R&D expenses of RMB10.8 million in 2023, RMB64.5 million in 2024 and RMB209.2 million in 2025, representing our largest operating cost center. Demonstrating emerging operating leverage, these expenses decreased meaningfully as a percentage of total revenue from 877.7% in 2024 to 378.1% in 2025. Our R&D expenditures primarily comprised (i) R&D computing resources costs, which represent computing power rental fees utilized for model adaptation, underlying operator library optimization, and platform testing, (ii) employee benefit expenses, which represent compensation, benefits, and share-based compensation for our research and development personnel, and (iii) other expenses, which primarily include professional service fees, depreciation and amortization, and traveling expenses. As the token supply industry remains fast-evolving and technology-intensive, continuous investment in R&D is critical for maintaining performance leadership, improving cost efficiency, expanding service capabilities and supporting future commercialization. Importantly, our R&D capabilities directly underpin our path to profitability: continuous technological breakthroughs enable us to improve inference capability, which is essential for structurally lowering our unit cost of revenue and improving our gross margins over time. We expect our absolute R&D expenses to increase as we continue to advance our token supply capabilities, while we anticipate these expenses will gradually decrease as a percentage of total revenue as we achieve greater commercial scale.

***Investments in sales and marketing and general and administrative expenses.*** Our profitability was further impacted by substantial expenditure in sales and marketing and general and administrative activities. Our sales and marketing expenses increased from RMB110 thousand in 2023 to RMB6.4 million in 2024, and further by 1,210.1% to RMB83.7 million in 2025, with promotional computing resources costs — representing the computing costs associated with the large-scale distribution of free token vouchers to acquire and onboard new users onto our public cloud platform — constituting the predominant driver at RMB54.2 million, or 64.7% of total sales and marketing expenses, in 2025. Distributing these vouchers served not only as a customer acquisition tool but also as a vital operational strategy: by generating massive real-world workloads, it helped us continuously refine our platform's usability, debug the underlying infrastructure, and optimize the overall user experience, directly contributing to the explosive growth of our registered user base to approximately 9.2 million by the end of 2025. As our platform and market presence mature, we plan to structurally reduce our reliance on widespread promotional voucher distributions, and accordingly anticipate that our sales and marketing expenses as a percentage of total revenue will steadily decrease, driving sustainable operating leverage. Our general and administrative expenses grew from RMB1.2 million in 2023 to RMB9.1 million in 2024, and further to RMB21.8 million in 2025, representing an increase of 140.2%, primarily attributable to employee benefit expenses as we expanded our administrative team to support our growing operations.

## Plan to Achieve Long-Term Profitability

We believe the sustainability of our business is supported by our ability to structurally improve gross margins and operational efficiency through the following measures:

### *Capitalizing on Explosive Industry Growth*

The token supply market is at an inflection point, with demand expanding at an unprecedented pace. We are well-positioned to capture a growing share of this expanding market, driven by the following factors:

- *Exponential growth in token demand.* According to Frost & Sullivan, China's token throughput surged from approximately 142.5 trillion tokens in 2024 to approximately 2,426.3 trillion tokens in 2025, representing year-on-year growth of 1,602.6%, and is projected to reach approximately 53.2 quintillion tokens by 2030, representing a CAGR of approximately 638.3% from 2025 to 2030. This exponential growth is underpinned by the rapid proliferation of AI applications across enterprises spanning coding, agents, multimodal generation. As AI applications continue to mature and token consumption per use case deepens, we expect demand for high-performance, cost-efficient token supply to grow significantly.

- *Leading market position to capture industry tailwinds.* As the largest independent ecosystem token supplier in China and a globally recognized token supply platform, we are well-positioned to capture a disproportionately large share of market growth as the industry scales. We expect robust industry tailwinds to underpin our long-term revenue and gross margin improvement trajectory.

***Deepening Domestic Customer Engagement***

We seek to grow domestic revenue by simultaneously expanding our customer base and deepening engagement with existing customers, through the following approaches:

- *Customer expansion and retention.* We serve customers across multiple segments and are particularly focused on deepening our penetration of high-value customers, who typically exhibit higher usage intensity, greater spending per account, and stronger long-term retention characteristics. Our high-value customer base has evolved materially in terms of usage scale during the Track Record Period. Customers with monthly token consumption over 100 billion began to emerge from the second half of 2025; customers consuming over 1 trillion tokens per month have appeared in 2026, and this highest-consumption cohort already accounts for more than half of, and an increasing share of, our total token throughput. As our penetration of higher-value segments increases and retention remains strong, we expect meaningful improvement in our average revenue per customer over time.

- *Migration to dedicated instance arrangements.* A key element of our enterprise strategy is the migration of customers from standard serverless token services – which we offer at competitive rates to drive platform adoption – to dedicated instances. Dedicated instances carry higher and more stable gross margins than standard serverless token services. As enterprise customers scale their AI usage on our platform, we expect a natural progression toward dedicated instances, which we believe will be a meaningful driver of gross margin expansion.

- *On-Premise deployment.* In addition to our public cloud-based services, our on-premise deployment business provides a high-margin complement to our growth strategy. Our on-premise deployment solutions generate gross margins of 82.5% in 2025, serving as a reliable source of profitability and cash generation that supports our overall financial position as we continue to invest in scaling our platform.

***Accelerating Overseas Revenue Growth***

We believe overseas markets represent a significant and relatively underpenetrated growth opportunity for our token supply services, and we plan to continue growing our overseas business through the following initiatives:

- *Expanding our overseas customer base through online and offline channels.* We have expanded our overseas presence by listing our serverless token services on a leading AI model routing platform. Building on this foundation, we plan to engage in more online and offline activities to increase brand visibility in key overseas markets.

- *Strengthening localized partnerships and building local operational capabilities.* We plan to recruit overseas sales, after-sales and technical personnel with relevant market experience to enhance customer service capabilities, market responsiveness and local delivery. We also intend to progressively establish and operate localized branch offices in selected overseas markets to support our local teams and strengthen our market presence in those jurisdictions.

***Maximizing Computing Resource Utilization***

We plan to maximize the productivity of our computing infrastructure through a combination of technology-driven improvements and intelligent scheduling strategies:

- *Heterogeneous computing adaptation*. We have developed the capability to adapt leading AI models for deployment on domestic chips, which carry substantially lower rental costs than NVIDIA GPU-based infrastructure. As the share of domestic AI chips within our computing infrastructure increases, we expect a material improvement in our overall cost structure. Our heterogeneous compatibility also reduces our dependence on any single computing resource provider, improving supply chain resilience and our ability to manage input cost volatility.

- *Inference engine optimization*. Our proprietary inference engine and underlying technology architecture are engineered to maximize computational throughput — enabling each chip to concurrently process a greater number of inference requests. This directly reduces our per-token production cost and underpins our ability to offer competitively priced services while expanding gross margin over time.

- *Off-Peak scheduling*. We are developing the capability to redirect inference capacity that is idle during off-peak hours toward batch inference workloads for academic institutions, research laboratories, and enterprise customers, offered at tiered pricing. This approach improves our around-the-clock cluster utilization rate and spreads our fixed computing resource costs across a larger volume of billable activity, directly compressing per-unit costs.

***Improving Operational Efficiency***

We plan to improve our cost structure and operational efficiency as our business scales through the following initiatives:

- *Procurement transformation*. During our early stage, we prioritized operational flexibility through short-term, on-demand computing resource rentals, which carried higher unit costs relative to longer-term arrangements. As our scale grows, we expect to develop greater bargaining power with computing resource providers, enabling us to progressively transition our procurement model from short-term on-demand rentals toward long-term annual contracts at materially lower unit rates, or toward joint operations arrangements under which suppliers bear a portion of the hardware costs.

- *Enhancing supply chain management*: We continue to strengthen our supply chain management capabilities by diversifying sources of computing resources and optimizing procurement and allocation, which supports operational stability and cost efficiency. We support a broad range of AI chip architectures, including both NVIDIA GPUs and domestic chips, and have developed capabilities to efficiently utilize heterogeneous computing resources across different architectures. This enhances our ability to secure computing resources supply, and improve cost efficiency.

Based on the foregoing, our Directors believe, and the Joint Sponsors concur, that our business is sustainable.

### Working Capital Sufficiency

For the year ended December 31, 2025, our monthly cash burn rate was RMB14.8 million. As of December 31, 2023, 2024 and 2025 we had cash and cash equivalents of RMB4.9 million, RMB49.7 million, and RMB171.5 million, respectively, and current term deposits of RMB nil, RMB nil, and RMB100.0 million, respectively. Our Directors are of the opinion that, taking into account the estimated net [**REDACTED**] from the [**REDACTED**] and other financial resources available to us, including cash and cash equivalents and current term deposits, we have sufficient working capital to cover at least 125% of our group's costs, including research and development expenses, sales and marketing expenses, and general and administrative expenses, for the next 12 months from the date of this document.

## OUR CORE TECHNOLOGIES

Our core technology foundation comprises our proprietary inference engine and computing resource orchestration system, which together form an integrated technology stack purpose-built for AI inference workloads. Our technology supports scalable and reliable AI inference across diverse models, computing architectures and deployment environments, enabling the efficient orchestration of heterogeneous computing resources and the transformation of underlying computing power into high-throughput, low-latency and cost-efficient token production.

We have established a comprehensive AI infrastructure technology system encompassing smart routing and observability of model APIs, high-performance inference acceleration for enterprise-grade models, and underlying scheduling. This enables our AI computing infrastructure to operate with high performance, high reliability, and high scalability in large-scale production environments. Our core technologies cover multiple key areas, aiming to address the computing power bottlenecks in the commercialization of models and providing customers with a low-latency, high-concurrency, and cost-effective computing foundation.

### High-Performance Model Inference Engine Technology

At the underlying computing power and computing equipment level, we focus on reducing the dependency on a single hardware ecosystem and have designed several core features for heterogeneous computing scenarios:

- ***Deep Adaptation and Optimization of Heterogeneous Computing Resources:*** We have achieved deep compatibility with mainstream international advanced chips, as well as mainstream domestic chips such as Ascend, MetaX, Moore Threads, Hygon, and T-Head. Through operator optimization, memory management optimization, communication and computation co-optimization, scheduling and sampling optimization, and parallel computing algorithms, we have effectively improved the actual computational efficiency of domestic hardware. This equips them with capabilities comparable to international advanced computing power in terms of throughput and inference accuracy, thereby significantly enhancing the overall effectiveness of heterogeneous AI computing clusters.

- ***Diversified Parallelism Strategies:*** At the architecture level, we employ EP and PD separation technologies to achieve cluster-level scheduling. This architecture supports ultra-high concurrent online requests and ultra-long context processing.

- ***Standardized Deployment Pipeline:*** We adopt a highly standardized and modular pipeline design, significantly shortening the deployment cycles and iteration time of AI models on complex heterogeneous hardware architectures.

### Computing Resource Orchestration Technology

To address common challenges in AI model workloads, such as computing resource idling, fragmentation, and multi-tenant isolation, we have developed proprietary orchestration technologies:

- ***Dynamic Resource Provisioning Paradigm:*** We are driving the upgrade of computing resources from the traditional "long-term reservation" model to a second-level dynamic provisioning model. The system can automatically and granularly allocate computing power based on real-time requests, load fluctuations, tenant needs, and model specifications.

- ***Cold Start Optimization and Security Isolation:*** Our core technology solves the cold start latency problem of deployments and supports seamless and smooth capacity scaling. Simultaneously, we have implemented a low-level multi-tenant security isolation mechanism tailored for heterogeneous computing environments. This maximizes the global utilization of computing resources while ensuring data security.

### Model API Smart Routing and Observability Technology

At the system software and network operations level, we have developed a traffic gateway and a refined operations center that support full-lifecycle management:

- ***Multi-Dimensional Dynamic Smart Routing:*** The platform provides unified model access and seamless protocol conversion interfaces, with a built-in advanced smart routing engine. The system dynamically schedules massive concurrent requests by comprehensively considering weight, priority, session affinity, real-time cluster load, context length, and specific business dimensions, ensuring that tasks are assigned to the optimal compute nodes.

- ***Full-Link Observability and Precise Metering:*** We have embedded unified observation and telemetry probes within the system to collect real-time performance and invocation data. At the same time, the platform possesses high-precision, multi-dimensional token metering capabilities, providing reliable data support for commercial settlement, cost accounting, and resource pool planning.

## RESEARCH AND DEVELOPMENT

Our ability to develop new technologies, design new services, and enhance existing services is critical for strengthening our market position. All of the R&D activities relating to our Specialist Technology Products were conducted in-house. We have full ownership of all intellectual property rights arising from such R&D activities. We do not rely on any material in-licensed third-party technologies for the development of our Specialist Technology Products. We did not outsource any R&D activities to third parties, nor did we engage in any collaboration with third parties that was material to the research and development of our Specialist Technology Products during the Track Record Period and up to the Latest Practicable Date. We have established interdisciplinary R&D capabilities that draw upon a diverse range of fields, such as computer science, applied mathematics, deep learning algorithms, distributed systems, low-level system optimization, AI framework engineering, and AI infrastructure development. During the Track Record Period, all of our research and development expenses were incurred for our Specialist Technology Products and amounted to RMB10.8 million, RMB64.5 million and RMB209.2 million in 2023, 2024 and 2025, respectively. As of the Latest Practicable Date, we are not aware of any pending or threatened legal claims, arbitral or administrative proceedings that may have a material influence on our research and development activities for any of our key Specialist Technology Products.

### Our Early Trading and R&D Activities

From our incorporation on August 29, 2023 through our first revenue generation in December 2023, we were primarily engaged in substantive R&D of our inference engine and related technologies. These activities constituted the core foundation of our trading operations during this period and were not mere preparatory activities. The following table sets out the key milestones and substantive activities undertaken during this period and their direct contribution to the development of our Specialist Technology Products:

| Period | Substantive R&D/Trading Activity | Contribution to Specialist Technology Products |
|---|---|---|
| **August 2023 to September 2023** | Our Group was incorporated on August 29, 2023. The founding team immediately commenced development of the inference engine and related technologies, including project initiation and architecture design. We incurred R&D expenses of RMB10.8 million and general and administrative expenses of RMB1.2 million in the period of August 29, 2023 to December 31, 2023. During the same period, our total R&D expenditure accounted for 89.1% of the annual total operating expenditure for the same period. | Established the core technical architecture of our token supply platform and foundational codebase |
| **October 2023** | Completed deployment and adaptation of multiple open-source models. Development efforts focused on model parallelism technologies and optimization of key inference engine components, including sampling optimization, scheduling optimization and memory management optimization, to improve inference efficiency and resource utilization. | Established and enhanced the core capabilities of our Specialist Technology Products |
| **November 2023** | Developed pipeline parallelism capabilities, expanded support for additional AI models and further enhanced the performance of inference engine. We also commenced the initial implementation of FP8 quantization and runtime optimization. | Enhanced the scalability and model compatibility of our Specialist Technology Products |
| **December 2023** | Commenced cross-chip architecture development, including the development of accelerator plug-ins to support deployment and optimization across heterogeneous computing architectures. | Established the foundation for cross-chip compatibility |
| **December 2023** | We started to generate revenue through software subscriptions, marking the beginning of our commercial operations. | First revenue generation validated the commercial viability of our Specialist Technology Products |

## Our R&D Process

Our R&D process encompasses six key stages — requirements analysis, project initiation, overall design, iterative development, quality verification, and deployment with ongoing operational monitoring and feedback — as set out below:

- *Requirements Analysis*. Each R&D initiative begins with a thorough analysis of market demand, customer requirements and technical feasibility. Our team assesses the commercial opportunity, evaluates alignment with our strategic priorities, and identifies the technical challenges and resource requirements involved. This stage establishes the core rationale for the project and defines the functional objectives that will guide all subsequent development work.

- *Project Initiation.* Once the requirements are validated, the R&D project is formally initiated through an internal approval process. Our R&D team prepares project documentation covering scope, technical direction, resource allocation, key milestones and success criteria. Through collaborative review and peer discussion, potential risks are identified and a clear, agreed-upon development roadmap is established before material resources are committed.

- *Overall Design.* With a confirmed project mandate, our engineers develop the high-level architecture and system design for the proposed product or technology. This includes defining the inference pipeline structure, API interfaces, scheduling logic, and hardware compatibility requirements. The overall design serves as the technical blueprint for iterative development and ensures alignment across the relevant R&D sub-teams.

- *Iterative Development.* Development proceeds through a series of focused iteration cycles. Our engineers implement and refine algorithms, conduct performance benchmarking across heterogeneous computing environments, and progressively integrate components into a functional system. Rapid prototyping and data-driven analysis are central to this stage, enabling the team to optimize throughput, reduce latency, and lower per-token production costs in an agile manner.

- *Quality Verification.* Prior to deployment, each product or technology release undergoes systematic quality verification. This includes functional testing, performance benchmarking against target SLA metrics, accuracy verification, security assessments and compatibility testing across the supported hardware architectures and model configurations. Only upon satisfactory completion of all verification requirements is a release approved for deployment.

- *Deployment, Operational Monitoring and Feedback.* Upon successful verification, the product or technology is deployed into production. Our operations team continuously monitors service availability, inference latency, throughput and system stability. Real-time telemetry and customer feedback are collected and analyzed, and insights from operational performance are fed back into the next development cycle to drive iterative improvements. This closed-loop mechanism ensures that our platform capabilities evolve in close alignment with the needs of our customers and the broader AI ecosystem.

## Our R&D Team and Core Members

As of the Latest Practicable Date, our in-house R&D department consisted of 67 personnel, a vast majority of whom hold a bachelor's degree or above. Our R&D department is organized into 6 core units, including foundation structure sector, public cloud sector, technical operation sector, design sector, innovative product sector and private cloud sector. Our R&D team is led by six core members, with details set forth in the following table:

| Core R&D Members | Profile |
|---|---|
| Dr. Yuan Jinhui | Dr. Yuan Jinhui received a bachelor's degree in computer science from Xidian University (西安電子科技大學) in 2003 and earned his Ph.D. from Tsinghua University (清華大學) in 2008, where his doctoral dissertation was recognized with an Outstanding Thesis Award under the supervision of Professor Zhang Bo (張鈸), who is a member of the Chinese Academy of Sciences (中國科學院) and the founding pioneer of AI research in China. Since then, Dr. Yuan has built a track record of translating frontier research into production-grade platforms and widely adopted innovations:<br>• Industrial AI Platform Engineering and Innovation (2013-2016): In 2013, Dr. Yuan joined Microsoft (China) Co., Ltd. (微軟(中國)有限公司). In 2014, he invented LightLDA, a high-efficiency topic model training algorithm.<br>• Distributed Deep Learning Platform (2017–2023): In 2017, Dr. Yuan founded OneFlow, a distributed deep learning platform project. At OneFlow, he led the development of the OneFlow deep learning framework, with a strategic focus on enhancing algorithm R&D efficiency and hardware utilization, helping to bridge research innovation with scalable computing performance.<br>• AI Infrastructure Entrepreneurship (2023–present): Since founding SiliconFlow in 2023, Dr. Yuan has led the Company's AI infrastructure initiatives. Dr. Yuan also actively engages with ecosystem partners and industry stakeholders, including participation in closed-door technical exchanges, to advance collaborative AI infrastructure ecosystems. |
| Mr. Liu Juncheng | Mr. Liu Juncheng is a principal development contributor to the Company's core technology stack, possessing strong capabilities in AI system software research and development. Mr. Liu holds a bachelor's degree in computer science and technology from Harbin Institute of Technology (哈爾濱工業大學). |
| Mr. Chen Houjiang | Mr. Chen Houjiang has over 10 years of experience in deep learning framework development, having previously served at Baidu Online Network Technology (Beijing) Co., Ltd (百度时代网络技术(北京)有限公司), Taobao (China) Software Co., Ltd. (淘宝(中国)软件有限公司) and OneFlow. Mr. Chen holds a bachelor's degree in communication engineering from China University of Geosciences (Wuhan) (中國地質大學(武漢)). |
| Mr. Liao Xingyu | Mr. Liao Xingyu specializes in deep learning algorithms, distributed training, and AI infrastructure perception, having previously served at AI Platform and Research Department of JD, OneFlow, and Beijing Huixi Intelligence Technology Co., Ltd. (北京輝羲智慧科技有限公司). Mr. Liao holds a master's degree in mathematics from the University of Science and Technology of China (中國科學技術大學). |

| Core R&D Members | Profile |
|---|---|
| Dr. Li Yipeng | Dr. Li Yipeng specializes in applied mathematics and distributed deep learning systems. Dr. Li currently holds 13 granted patents in China and the United States, with an additional 17 patents in preparation or under application. Dr. Li holds a bachelor's degree in information and computing science from the University of Science and Technology of China (中國科學技術大學) and a Ph.D. in applied mathematics from Stony Brook University in the United States. |
| Mr. Zhang Wenxiao | Mr. Zhang Wenxiao has over ten years of full-stack development experience spanning system development, AI frameworks, and deep learning infrastructure, and is a founding-level core contributor to the OneFlow deep learning framework. Mr. Zhang holds a bachelor's degree in software engineering from Sichuan University (四川大學) and a master's degree in software engineering from Huazhong University of Science and Technology (華中科技大學). |

The above core R&D members are the key personnel responsible for the technical operations and the research and development of our Specialist Technology Products, since they hold primary technical leadership responsibility for the development, optimization and ongoing advancement of our Specialist Technology Products.

To attract and retain top-tier management and technical talent, we offer competitive compensation packages and comprehensive incentive programs. We are also committed to the long-term development of our workforce by providing structured career development pathways that deepen technical expertise and encourage cross-disciplinary collaboration, while enabling employees to expand their skills through hands-on experience in complex projects.

To preserve institutional expertise and support business continuity, we maintain robust knowledge retention and documentation systems across our R&D, product and project management functions. These continuously updated repositories facilitate the effective transfer of critical technical knowledge within our teams, while remaining subject to our information security and confidentiality protocols. We also manage employee transitions in a professional and orderly manner. Through structured handover procedures and established documentation, access control and project backup mechanisms, we seek to ensure continuity of critical technical operations and minimize disruption to team collaboration.

The key terms of our agreements with significant R&D personnel are set out below:

- *Ownership of the intellectual property:* Any inventions, creations, works, or non-patented technical achievements created by employees during their employment as a result of performing their duties or primarily utilizing our materials and technological conditions, and business information, among other things, including trade secrets, shall belong to us. We have the right to use or transfer the aforementioned intellectual property. Employees shall actively provide all necessary information, materials, and research, and take all necessary actions to assist us in obtaining and exercising the relevant intellectual property rights.

- *Confidentiality:* Employees undertake not to disclose or use any confidential information.

- *Non-competition and non-solicitation:* We reserve the right to unilaterally impose a non-competition period of up to two years following the termination of employment, with appropriate compensation provided accordingly. During the term of employment and the non-competition period imposed by us, the employee shall not engage in any competitive behavior, and they shall not solicit or attempt to solicit our employees to leave their employment.

- *Dispute resolution:* The parties shall resolve any issues arising through negotiation. If such issues remain unresolved, parties may seek arbitration from the labor arbitration committee at our location.

## INTELLECTUAL PROPERTY

Intellectual property rights are crucial to our business. Our future commercial success depends, in part, on our ability to obtain and maintain patents and other intellectual property rights and proprietary protections for commercially important technologies, inventions and know-how related to our business, defend and enforce our patents, preserve the confidentiality of our trade secrets, and operate without infringing, misappropriating or otherwise violating the intellectual property rights of third parties.

As of the Latest Practicable Date, we owned 11 registered patents and 34 patent applications, 17 software copyrights, and six registered trademarks in China. As of the same date, we also owned one registered trademark in Hong Kong, two in Singapore, one in Japan and one in the United States. As of the Latest Practicable Date, we owned all our patents as well as patent applications and had no co-ownership or co-sharing arrangements of our patents and patent applications with third parties.

For material intellectual property rights in relation to each Specialist Technology Product, see "Appendix IV — Statutory and General Information — B. Further Information about Our Business — 2. Intellectual Property Rights" for details on the material patents for our core technologies in relation to our Specialist Technology Products. During the Track Record Period and up to the Latest Practicable Date, we had not been involved in any material legal, arbitral, or administrative proceedings or claims of infringement of any intellectual property rights, in which we may be a claimant or a respondent. Our Directors confirm that they are not aware of any material legal, arbitral or administrative proceedings for infringement of any third party's intellectual property rights by us as of the Latest Practicable Date.

## SALES AND MARKETING

### Sales and Marketing Team

As of December 31, 2025, we had established a sales and marketing team consisting of 21 personnel with coverage across both the PRC and overseas markets. Our sales team is organized according to business segment and customer tier to enhance customer reach and responsiveness, with dedicated coverage for public cloud customers and on-premise enterprise clients respectively. Our business development and sales teams work closely with our research and development team to ensure that technical capabilities can be effectively communicated to customers and translated into commercial engagements.

We primarily sell our products and services through direct sales and channel partnerships. Given the nature of our business – which involves delivering token services over the cloud or through software licensing – our commercial model does not involve physical product distribution or traditional reseller channels. We have also expanded our overseas presence by listing our serverless token services on a leading AI model routing platform.

### After-Sales Services

We consider after-sales service an important component of customer experience and platform reliability management. We have established a dedicated after-sales and operations team responsible for monitoring platform stability, responding to customer service requests, and managing incident resolution across both public cloud and on-premise deployments.

For public cloud customers, we maintain continuous real-time monitoring of platform availability, latency, and throughput performance. When anomalies are detected, our operations team is alerted immediately and works to identify root causes and restore normal service levels. For on-premise deployment customers, we manage after-sales engagement through a structured service protocol. Customers contact our service team through designated channels upon encountering post-deployment issues, and each case is tracked through our internal service management system. Our commitments include responding to incidents within one to four hours and resolving within 12 to 48 hours, depending on the severity of the incidents. We maintain a dedicated team of delivery and support engineers who provide both remote and on-site assistance as required.

During the Track Record Period, we did not experience any customer contract terminations or major customer complaints.

## Marketing

Our marketing strategy is designed to continuously strengthen brand recognition, broaden our reach among developers and enterprise customers, drive product adoption and customer conversion, and further consolidate our leading position in the token supply sector. Anchored in our independent ecosystem strategy, we have established a multi-tiered market development framework spanning developers, enterprise customers, and industry partners, and we pursue sustained business growth through brand building, ecosystem operations, product promotion, and commercial development initiatives.

At the developer ecosystem level, we continuously expand our user reach and enhance product awareness and developer engagement through developer events, content marketing, and media outreach. By offering a broad selection of models, cost-effective inference services, and a streamlined development experience, we attract developers and innovative teams to adopt our products and services, driving growth in our user base and token volume.

At the enterprise customer level, we showcase our technical capabilities, product strengths, and industry use cases through industry events, technology summits, customer engagement activities, and solution promotions, thereby facilitating enterprise customers' understanding and adoption of our products and services. For customers with commercial potential, our marketing, business development, and sales teams collaborate closely to conduct needs assessments and match appropriate solutions.

We also actively engage in ecosystem collaboration with model providers, chip designers, cloud service providers, and other industry partners. Through joint promotions, technical collaboration, and co-development of ecosystem initiatives, we seek to expand market opportunities, enhance our ecosystem influence, and strengthen the overall value of our platform.

## CUSTOMERS

In 2023, 2024 and 2025, our aggregate revenue from the five largest customers in each year during the Track Record Period was RMB6 thousand, RMB6.2 million and RMB24.9 million, respectively, accounting for 100.0%, 85.0% and 45.0% of our total revenue for the respective years. In the same years, our revenue from the single largest customer in each year during the Track Record Period was RMB5 thousand, RMB4.5 million and RMB7.5 million, accounting for 83.3%, 61.1% and 13.6% of our total revenue for the respective years.

The following tables set forth the details of our top five customers during the Track Record Period:

| Customers | Products/services | Year of commencement of business relationship | Credit period | Payment method | Revenue | Percentage of total revenue |
|---|---|---|---|---|---|---|
| | | | | | *(RMB'000)* | |
| **For the year ended December 31, 2023** | | | | | | |
| Customer A [(1)] | On-premise deployment solutions | 2023 | Pay upon order | Bank transfer | 5 | 83.3% |
| Customer B [(2)] | On-premise deployment solutions | 2023 | Pay upon order | Bank transfer | 1 | 16.7% |
| **Total** | | | | | **6** | **100.0%** |

| Customers | Products/services | Year of commencement of business relationship | Credit period | Payment method | Revenue | Percentage of total revenue |
|---|---|---|---|---|---|---|
| | | | | | *(RMB'000)* | |
| **For the year ended December 31, 2024** | | | | | | |
| Customer C[3] | On-premise deployment solutions | 2024 | 5 or 10 working days | Bank transfer | 4,487 | 61.1% |
| Customer D[4] | On-premise deployment solutions | 2024 | Prepayment | Bank transfer | 1,040 | 14.2% |
| Customer E[5] | Dedicated instances | 2024 | Prepayment | Bank transfer | 377 | 5.1% |
| Customer F[6] | On-premise deployment solutions | 2024 | Prepayment | Bank transfer | 175 | 2.4% |
| Customer G[7] | On-premise deployment solutions | 2024 | 5 working days | Bank transfer | 164 | 2.2% |
| **Total** | | | | | **6,243** | **85.0%** |

| Customers | Products/services | Year of commencement of business relationship | Credit period | Payment method | Revenue | Percentage of total revenue |
|---|---|---|---|---|---|---|
| | | | | | *(RMB'000)* | |
| **For the year ended December 31, 2025** | | | | | | |
| Customer H[8] | On-premise deployment solutions | 2025 | 5 working days | Bank transfer | 7,547 | 13.6% |
| Customer I[9] | On-premise deployment solutions | 2025 | 30 days | Bank transfer | 7,100 | 12.8% |
| Customer J[10] | Dedicated instances | 2025 | 30 days or prepayment | Bank transfer | 4,138 | 7.5% |
| Customer K[11] | On-premise deployment solutions | 2024 | 14 working days | Bank transfer | 3,189 | 5.8% |
| Customer L[12] | On-premise deployment solutions | 2025 | 15 working days | Bank transfer | 2,903 | 5.2% |
| **Total** | | | | | **24,877** | **45.0%** |

*Notes:*

(1) Customer A is an individual whose email address is associated with a private generative AI software company.

(2) Customer B is an individual whose email address is associated with an education platform.

(3) Customer C is a company headquartered in China, focusing on information system integration services and artificial intelligence basic resources and technology platforms, artificial intelligence industry application system integration services, integrated circuit design and related technology services.

(4) Customer D is a company headquartered in China, focusing on technology development, software development, computer systems services, information systems integration and related technology services.

(5) Customer E is a company headquartered in China, focusing on information systems integration and AI industry application systems integration.

(6) Customer F is a company headquartered in China, focusing on software and information technology services.

(7) Customer G is a company headquartered in China, focusing on information system integration services and technology promotion.

(8) Customer H is a company headquartered in China, focusing on the R&D, design and sales of full-stack chip products and related software stacks and computing platforms for AI training and inference, general computing and graphics rendering.

(9) Customer I is a company headquartered in China and its affiliates, focusing on information and communications technology infrastructure, smart devices, cloud computing and related ICT solutions, products and services.

(10) Customer J is a company headquartered in China, focusing on technology services, software development, computer systems services and related computer software and hardware business.

(11) Customer K is a research institute based in China, focusing on original scientific and technological innovation research.

(12) Customer L is a company headquartered in China, focusing on information system integration, operation and maintenance (O&M) software development and data processing.

To the best of our knowledge, all of our five largest customers in each year during the Track Record Period were independent third parties. As of the date of this document, none of our Directors, their associates or any of our Shareholders (who or which to the knowledge of the Directors owned more than 5% of our issued share capital) had any interest in any of our five largest customers in each year during the Track Record Period.

**Key Terms of Serverless Token Service Contracts**

Key terms of our serverless token service contracts with our customers include:

- *Services*. We generally provide serverless token service to the customers, and offer technical support, including application deployment and program debugging.

- *Payment methods*. The customers generally purchases our serverless service by (i) sending emails and transferring the fees to us, or (ii) topping up their accounts on the platform. We may allow the customers to use the service prior to payment in certain cases.

- *Service limits*. The customers enjoy exclusive and non-transferable rights to use the service for personal use or the use of entities represented by them, subject to prohibitions from, among others, (i) reverse engineering, decoding or decompilation, (ii) transferring or sublicensing API keys without our prior written consent, (iii) breach of our intellectual property, (iv) using our service to process, transmit, or store unlawful content or information, and (v) breaking or undermining our servers or infrastructures. We reserve the rights to process the interaction data from the service to the customers in accordance with applicable laws, provided we determine that such data violate laws, regulations or this contract.

- *Intellectual property*. The customers understand and acknowledge that we continue to enjoy all the intellectual properties under the service and is prohibited from using such intellectual properties for any purpose unless the contract permits otherwise. The customers may use the generated outputs, provided that such use does not transfer, infringe or otherwise violate any intellectual property rights of ours or any third party.

- *Confidentiality*. The customers are required not to disclose any confidential information of ours or of other users accessible through the service.

- *Termination*. Each party is generally entitled to terminate the contract by giving the notice period specified in the relevant contract in advance.

**Key Terms of Dedicated Instance Contracts**

Key terms of our dedicated instance contracts with our customers include:

- *Services*. We generally provide dedicated instances to the customers, and offer technical support, including application deployment and program debugging. For certain customers, we also provide an SLA guaranteeing minimum standards for latency, throughput and stability.

- *Payment method*. The customers either pay (i) a fixed subscription fee for the reserved period, or (ii) usage-based charges measured in tokens consumed within the exclusive instance.

- *Service limits*. The customer enjoy exclusive and non-transferable rights to use the service for personal use or the use of entities represented by them, subject to prohibitions from, among others, (i) reverse engineering, decoding or decompilation, (ii) transferring or sublicensing API keys without our prior written consent, (iii) breach of our intellectual property, (iv) using our service to process, transmit, or store unlawful content or information, and (v) breaking or undermining our servers or infrastructures. We reserve the rights to process the interaction data from the service to the customers in accordance with applicable laws, provided we determine that such data violate laws, regulations or this contract.

- *Intellectual property*. The customers understand and acknowledge that we continue to enjoy all the intellectual properties under the service and is prohibited from using such intellectual properties for any purpose unless the contract permits otherwise. The customers can use the generated outcomes whereto there will be no transfer or breach of any of our or third parties' intellectual properties.

- *Confidentiality*. The customers are required not to disclose any confidential information of ours or of other users accessible through the service.

- *Extension*. The contract is generally automatically extended unless either party terminates it by giving the notice period specified in the relevant contract prior to the expiration date.

**Key Terms of On-premise Deployment Solutions Contracts**

Key terms of our on-premise deployment solutions contracts with our customers include:

- *Services*. We generally deploy our inference engine and computing resource orchestration system within the customers' own data center or private computing environment, with ongoing maintenance and support provided following deployment.

- *Payment*. Fees are generally structured as a combination of a software licensing fee (on a subscription or one-time basis), a professional implementation fee and, where applicable, a maintenance fee.

- *Product and service limits*. The customers should obey applicable laws when using the product and service. A breach of relevant terms of the contract by the customers will entitle us to cancel the authorization without bearing liabilities.

- *Intellectual property*. We own the intellectual properties of the software. The customers are prohibited from reverse engineering, decoding or decompilation, and are forbidden to use such intellectual properties for any purpose unless the law permits or we allow otherwise.

- *Confidentiality*. Both parties agree to use confidential information of each other in strict compliance with the contract and are prohibited from disclosing it to third parties unless the other party gives written consent in advance.

- *Termination*. The two parties may mutually agree to terminate the contract. The contract is generally automatically extended unless either party terminates it by giving the notice period specified in the relevant contract prior to the expiration date. We are also entitled to unilaterally terminate the contract in writing by giving the notice period specified in the relevant contract in advance.

## SUPPLIERS

In 2023, 2024 and 2025, our aggregate purchases from the five largest suppliers in each year during the Track Record Period was RMB2.2 million, RMB11.4 million and RMB117.3 million, respectively, accounting for 100.0%, 87.6% and 70.8% of our total purchases for the respective years. For the same years, our purchase from the single largest supplier in each year during the Track Record Period amounted to RMB2.2 million, RMB6.8 million and RMB33.8 million, accounting for 100.0%, 52.1% and 20.4% of our total purchase for the respective years.

The following tables set forth the details of our top five suppliers during the Track Record Period:

| Suppliers | Products/services provided | Year of commencement of business relationship | Credit period | Payment method | Purchase | Percentage of total purchase |
|---|---|---|---|---|---|---|
| | | | | | *(RMB'000)* | |
| **For the year ended December 31, 2023** | | | | | | |
| Supplier A[(1)] | Computing resources | 2023 | N/A | Bank transfer | 2,169 | 100.0% |
| **Total** | | | | | **2,169** | **100.0%** |

| Suppliers | Products/services provided | Year of commencement of business relationship | Credit period | Payment method | Purchase | Percentage of total purchase |
|---|---|---|---|---|---|---|
| | | | | | *(RMB'000)* | |
| **For the year ended December 31, 2024** | | | | | | |
| Supplier A[(1)] | Computing resources | 2023 | 10 working days | Bank transfer | 6,774 | 52.1% |
| Supplier B[(2)] | Computing resources | 2024 | Prepayment and 10 days | Bank transfer | 2,206 | 17.0% |
| Supplier C[(3)] | Computing resources | 2024 | Prepayment | Bank transfer | 1,091 | 8.4% |
| Supplier D[(4)] | Computing resources | 2024 | Prepayment and 10 days | Bank transfer | 700 | 5.4% |
| Supplier E[(5)] | Cloud services | 2024 | Prepayment | Bank transfer | 604 | 4.6% |
| **Total** | | | | | **11,375** | **87.6%** |

| Suppliers | Products/services provided | Year of commencement of business relationship | Credit period | Payment method | Purchase | Percentage of total purchase |
|---|---|---|---|---|---|---|
| | | | | | *(RMB'000)* | |
| **For the year ended December 31, 2025** | | | | | | |
| Supplier F[6] | Computing resources | 2025 | Prepayment | Bank transfer | 33,830 | 20.4% |
| Supplier G[7] | Computing resources | 2025 | Prepayment | Bank transfer | 25,328 | 15.3% |
| Supplier H[8] | Computing resources | 2025 | Monthly settlement | Bank transfer | 24,491 | 14.8% |
| Supplier I[9] | Computing resources | 2024 | Prepayment | Bank transfer | 21,005 | 12.7% |
| Supplier J[10] | Computing resources | 2025 | Prepayment | Bank transfer | 12,616 | 7.6% |
| **Total** | | | | | **117,270** | **70.8%** |

*Notes:*

(1) Supplier A is a company headquartered in China, focusing on deep learning frameworks, AI operating systems, software development and related technical services.

(2) Supplier B is a company headquartered in China, focusing on computing power services, cloud services, internet data services, software development, information systems integration and related technology services.

(3) Supplier C is a company headquartered in China, providing cloud computing, IDC hosting, IT value-added, IT outsourcing and communications integration services to customers across industries.

(4) Supplier D is a company headquartered in China, focusing on network technology and related cloud/IT services.

(5) Supplier E is a company headquartered in China, focusing on cloud computing services.

(6) Supplier F is a company headquartered in China, focusing on full-stack cloud services, including public cloud, private cloud, dedicated cloud, hybrid cloud and edge cloud services.

(7) Supplier G is a company headquartered in China, focusing on technology services, software development, computer systems services, artificial intelligence application software development, artificial intelligence basic software development and information systems integration.

(8) Supplier Group H refers to Supplier H and its affiliates. Supplier H is a telecommunications company headquartered in China, providing telecommunications services.

(9) Supplier I is a company headquartered in China, focusing on cloud computing and related technology services.

(10) Supplier J is a company headquartered in China and its affiliates, focusing on information and communications technology infrastructure, smart devices, cloud computing and related ICT solutions, products and services.

To the best of our knowledge, except for Supplier A and Supplier I, all of our five largest suppliers in each year during the Track Record Period were independent third parties. As of the Latest Practicable Date, except for Supplier A and Supplier I, none of our Directors, their associates or any of our Shareholders (who or which to the knowledge of the Directors owned more than 5% of our issued share capital) had any interest in any of our top five suppliers in each year during the Track Record Period.

Key terms of our agreements with our suppliers include:

- *Services*. Our suppliers generally provide computing resources and related technical support services, with the service scope, configuration, period and fees set out in the relevant order, service list or platform records.

- *Payment*. We generally settle supplier fees through prepayment or monthly payment arrangements. Fees are typically calculated based on fixed resource packages, actual usage, bandwidth usage or supplier platform records.

- *Service support*. Suppliers generally provide operation and maintenance services, technical support and service availability commitments.

- *Compliance and data protection*. We are generally required to use the suppliers' services in compliance with applicable laws, regulatory requirements, platform rules and network security policies. Suppliers are generally required to keep our business data and confidential information confidential.

- *Term*. Supplier agreements generally have fixed service terms or framework terms. Early termination may require prior written notice.

## OVERLAPPING CUSTOMERS AND SUPPLIERS

During the Track Record Period, to the best knowledge and belief of our Directors, two of our major business partners served as both customers and suppliers to us.

Customer I, which is also Supplier J, is our second largest customer and our fifth largest supplier in 2025. We mainly provide on-premise deployment solutions to Customer I, and Supplier J mainly provides computing resources to us through a different entity within the group. In 2023, 2024 and 2025, our sales to Customer I amounted to nil, nil and RMB7.1 million, accounting for nil, nil and 12.8% of our total revenue, respectively. During the same periods, our purchases from Supplier J amounted to nil, nil and RMB12.6 million, accounting for nil, nil and 7.6% of our total purchase amount, respectively.

Supplier I is our fourth largest supplier in 2025 and is also our customer. We provided serverless token services to Supplier I in 2025 on an incidental basis, and Supplier I mainly provides computing resources to us. In 2023, 2024 and 2025, our sales to Supplier I amounted to nil, nil and RMB8, accounting for nil, nil and less than 0.1% of our total revenue, respectively. During the same periods, our purchases from Supplier I amounted to nil, RMB419 thousand and RMB21.0 million, accounting for nil, 3.2% and 12.7% of our total purchase amount, respectively.

According to Frost & Sullivan, it is common in the token supply industry for customers and suppliers to overlap. Our Directors confirmed that all transactions with the above overlapping customers and suppliers were conducted on an individual basis and that the relevant sales and purchases were neither interconnected nor conditional upon each other. All such transactions were entered into in the ordinary course of business under normal commercial terms and on an arm's length basis.

## DATA SECURITY AND PRIVACY

A core principle of our platform design is that we do not store or use customer interaction data – including prompts submitted to and outputs generated by the models available on our platform.

In the ordinary course of our business, we collect and maintain limited personal information from our users, which comprises: (i) registration information, such as name and contact details, provided upon account creation; and (ii) identity verification and payment information, required by actual business operations. This personal information is stored separately from interaction data and is processed in accordance with applicable laws and regulations on data privacy and security in China and, in respect of our international platform, in the jurisdictions in which we operate. Although we operate both a domestic and an international platform, customers' personal data from each platform is stored separately. Interaction data is randomly scheduled based on computing resource availability and is deleted immediately upon completion. We do not provide personal data of our users to third parties except where required by applicable law or with the user's prior consent.

For customers using our on-premise deployment solution, all communications and inference requests are stored and processed within the customer's own infrastructure, with no exposure to our shared cloud environment.

We have implemented a range of technical and organizational measures to protect the security of data processed on our platform, including:

- Access control. We maintain an access control system for personal information in our internal systems, restricting access to authorized personnel only. Employee access to data is strictly limited based on role and function.

- Network and system security. We maintain firewalls and network security protocols to prevent information loss or leakage resulting from cyber-attacks, and we conduct periodic data security risk assessments.

- Compute isolation. Our cloud infrastructure implements computing isolation and network isolation across tenants to prevent unauthorized access to or leakage of customer data across the multi-tenant environment. We continue to evaluate and strengthen our isolation architecture, including exploring enhancements to container-level security.

- Employee confidentiality obligations. All employees are required to sign confidentiality and intellectual property assignment agreements upon joining the Company, covering confidentiality, non-compete obligations, and the treatment of intellectual property created during employment. Employees involved in important customer projects are additionally required to execute project-specific confidentiality agreements.

- Data compliance governance. We have established a data compliance system covering each department's data processing methodology, emergency response and contingency plans for information security incidents, and an internal working group that conducts regular monitoring of data compliance matters.

During the Track Record Period and up to the Latest Practicable Date, we did not encounter any material malfunction, unexpected system failure, interruption, or security breach in our information technology and software systems, nor did we experience any material data leakage, data loss, or unauthorized use of customer personal information. Furthermore, we have not been subject to any litigations, arbitrations, or material administrative penalties in relation to data protection or cybersecurity that resulted in a material adverse effect during the same periods. We believe that, as a result of our strict internal controls, compliance governance, and our fundamental design principle of not storing or using customer interaction data, our business operations are in compliance with current data security laws and regulations in all markets where we operate in all material respects.

## ENVIRONMENTAL, SOCIAL, AND GOVERNANCE MATTERS

### ESG Governance

We strictly comply with national and local laws, regulations, and regulatory documents concerning artificial intelligence, ecological and environmental protection, labor and employment, and business ethics, and steadily enhance our ESG management performance. The Board of Directors serves as our highest supervisory and decision-making body for ESG governance. It is primarily responsible for reviewing major ESG-related matters, overseeing ESG practices, and monitoring the progress of our ESG targets. In our operations and major decision-making, we assess ESG risks and opportunities, drive the integration of sustainability principles and responsible business practices across the entire value chain through a top-down approach, and adhere to the highest standards of business ethics. The Board of Directors oversees ESG initiatives on an ongoing basis through regular meetings and dedicated

issue-based communications. In addition, we conduct ESG training at least annually to enhance the Board's knowledge of ESG regulatory frameworks and industry best practices, and to continuously strengthen the Board's professional competence in sustainable governance.

To better advance our ESG initiatives, we have established an ESG Working Group comprising management personnel from various departments. Its key responsibilities include: (i) identifying ESG risks and opportunities and formulating risk mitigation measures, (ii) developing ESG-related policies and tracking their implementation, (iii) assessing the impacts of climate change and formulating coping strategies, (iv) coordinating ESG data management and daily ESG administrative workflows, and (v) organizing ESG-related training for employees across all departments.

We are currently formulating our ESG policies with reference to the relevant requirements of the Listing Rules of the Stock Exchange of Hong Kong (《上市規則》). Post-[**REDACTED**], we will establish a normalized ESG information disclosure system, and regularly publish ESG Reports to practically safeguard the rights and interests of our stakeholders.

**ESG Risk Management**

The Directors, having made all reasonable inquiries, confirm that as of the Latest Practicable Date, we are in full compliance with all applicable ESG laws and regulations across all operating jurisdictions, with no material non-compliance affecting our operations, financials, performance, or reputation, nor any material regulatory penalties received.

To continuously enhance the effectiveness of risk management and control, the ESG Working Group is responsible for specific ESG risk assessments, root cause analysis, and the formulation of corresponding emergency response mechanisms, with relevant outcomes submitted to the Board of Directors at least once a year. Moving forward, we will continuously improve our ESG risk control system to build a solid foundation for sustainable corporate development.

**Materiality Assessment**

We construct our ESG material issue pool based on the ESG regulatory guidance documents of the Stock Exchange of Hong Kong and the Sustainability Accounting Standards Board Standards, while integrating our actual conditions. We distribute questionnaires to stakeholders, including our Directors, senior management, suppliers, and other relevant parties, and determine the materiality ranking of ESG issues through a combination of expert judgment and collective deliberation by the Board of Directors. According to the assessment results, we consider issues such as product responsibility, R&D and innovation, intellectual property protection, supply chain management, development and training, labor standards, anti-corruption, and climate change to be the key ESG issues that require our focused attention, and we place special emphasis on these aforementioned issues when implementing ESG management, control, and practices.

**Environmental**

Based on the nature of our business, we do not engage in manufacturing activities, and our environmental risks are relatively low. We formulate internal Environmental Protection Policies and advocate for green and low-carbon concepts. We promote green offices, standardize the use and management of electrical equipment, and encourage employees to adopt green travel practices. As of the Latest Practicable Date, we have obtained the ISO 14001:2015 Environmental Management System Certification. During the Track Record Period, we did not experience any environment-related complaints or penalties, and our cumulative expenses incurred in connection with environmental compliance during the Track Record Period was approximately RMB40 thousand.

***Environmental Management Targets***

To align with China's national Dual Carbon strategy, using 2025 as the baseline year, we have committed to reducing energy consumption intensity by 5% by 2028 and 10% by 2030, and reducing greenhouse gas emission intensity by 5% by 2028 and 10% by 2030. To ensure the implementation of these targets, we will implement these targets in phases through measures such as energy efficiency improvements, clean energy substitution, and the enhancement of employee environmental awareness.

***Waste Management***

Our daily business operations do not involve the discharge of hazardous waste and only generate non-hazardous waste, which primarily consists of domestic office waste generated from employees' daily office activities. The non-hazardous waste is centrally collected and disposed of by the property management of the building where our offices are located. Meanwhile, we encourage methods such as paperless office practices to reduce waste generation.

***Energy Management and Water Resource Utilization***

We strictly comply with laws and regulations relating to energy management and continuously promote the efficient utilization of resources. The energy consumed in our daily operations includes purchased electricity and office water. Given our business nature, almost all of our computing power is acquired through market-based leasing. As our business continues to expand, both headcount and office premises have been scaling up since 2024, which has directly led to a corresponding increase in the scale of energy consumption. To further practice the concept of green operations, we implement energy-saving management and control measures among all staff on a normalized basis, guiding employees to properly implement water and electricity conservation requirements, thereby continuously reducing resource consumption from the operational end. During the Track Record Period, our energy consumption is as follows:

| Indicator | Unit | 2023 | 2024 | 2025 |
|---|---|---|---|---|
| Total Electricity Consumption . . | kWh | 1,162.50 | 9,432.49 | 30,352.00 |
| Electricity Intensity . . . . . | kWh/RMB'000 of revenue | 193.75 | 1.28 | 0.55 |

(1) Since the commencement of our operations, all water-related expenses for our office premises have been consistently covered by the property management service provider of the building. Given that the office building has not installed separate meters to measure our actual water consumption, data regarding water resource usage is currently unavailable.

In addition to daily operations, we integrate green concepts into the full lifecycle management of products, spanning from technical R&D to commercial application. Through a combination of self-developed technologies, including a high-performance inference engine, lossless precision dynamic quantization, multi-level caching and smart routing, and the unified scheduling of heterogeneous computing resources, we significantly reduce computing resource consumption during the token production process, thereby directly lowering energy consumption. For deployed models including MiniMax-M2.5 and DeepSeek-V3.2, our technical stack delivers a 60%—80% reduction in inference compute volume on reserved instances (dedicated compute) versus the unoptimized baseline. Under actual system-level benchmarking, MiniMax-M2.5 registers a per-token energy footprint of 49.2 Wh per million tokens (at datacenter level with a power usage effectiveness of 1.3), which is approximately 39% lower than the Dense model reference point audited under MLPerf Inference v4.1 (Llama2-70B at 80.7 Wh per million tokens). On a parameter-normalized basis, this equates to a roughly 10.7-fold gain in energy efficiency. In cache-hit scenarios, compute and energy savings exceed approximately 90%. Additionally, we have consolidated chips previously constrained by software ecosystem limitations into a centrally managed compute scheduling pool, substantially improving overall compute resource utilization and curtailing energy losses attributable to idle capacity.

For commercial application scenarios, we leverage the elastic scheduling capabilities of cloud-native software to automatically match optimal resources for AI model inference tasks, and promptly release resources down to zero for models that do not need to run, avoiding the generation of invalid computing power and corresponding energy loss within data centers. In addition, relying on the cross-regional layout of computing power resources, we dynamically allocate computing power in light of local energy endowments and customers' business requirements. Computing resources are prioritized for deployment in regions rich in photovoltaic power during daytime and shifted to sites with abundant wind power at night to raise the proportion of renewable electricity consumption. Moving forward, we will continuously and dynamically optimize our computing power scheduling strategies, giving priority to utilizing computing power resources in regions with abundant green electricity such as solar and wind energy.

***Addressing Climate Change***

We pay close attention to various potential impacts triggered by climate change, and prudently assess and judge various risks and uncertainties brought by climate factors to our own operations, business development, and financial performance. To systematically manage climate-related risks, our General Administration Department leads the annual identification and assessment of climate risks and opportunities, covering: (i) physical risks — primarily the physical impacts of climate change, including acute risks from extreme weather events and chronic risks arising from global warming; and (ii) transition risks — primarily the impacts of transitioning to a low-carbon economy, including legal and policy risks, market risks, technological risks, and reputational risks. The assessment also analyzes the transmission effects of different climate risks on our daily operations as well as medium- and long-term development. In response to the identified climate risks with high impact levels, we continuously enhance our risk management capabilities and steadily strengthen our climate resilience and operational robustness.

As of the Latest Practicable Date, we have not incurred any material adverse losses to our operations or financial condition arising from climate change-related risks. The table below presents our identified climate-related risks, the impacts, and the measures taken or planned in response:

| Risk Description | Degree of Impact | Impact Period | Potential Impact | Mitigating Measures |
|---|---|---|---|---|
| Extreme weather may disrupt employee commuting and suspend office operations, posing potential risks to our stable operation. . . . . | Low | Short-term | Increased Operating Costs | Equip sufficient emergency response supplies, implement normalized potential safety hazard inspections of venues, actively conduct emergency drills, and improve off-site remote emergency office contingency plans. |
| Global warming may lead to higher energy consumption for us. . . . . | Low | Medium-term | Increased Operating Costs | Adopt energy-saving equipment and establish an energy consumption inspection mechanism, with normalized energy consumption management and control to tap deeper into the potential for consumption reduction. |

| Risk Description | Degree of Impact | Impact Period | Potential Impact | Mitigating Measures |
|---|---|---|---|---|
| Against evolving Dual Carbon Goals and regulatory caps on carbon emissions & energy consumption across the computing sector, non-compliant carbon management by our computing partners may restrict relevant cooperation. . . . . | Medium | Medium-term | Rising Procurement Costs for Computing Power Leasing | Track relevant laws and regulations on a normalized basis, monitor greenhouse gas emissions of upstream computing partners and enhance full-chain climate compliance management. |
| Extreme weather and energy volatility may push up power prices, with upstream computing providers passing higher power costs to rental quotes. Constrained by ESG management, downstream clients demand more renewable-powered computing. . . . . | Medium | Medium-term | Rising Procurement Costs for Computing Power Leasing | Expand qualified low-carbon computing resources powered by renewable energy; cooperate with partners to boost green power substitution for computing load. |
| If our inference energy efficiency ratio fails to continuously improve, this may potentially diminish our core competitiveness. . . . . | High | Long-term | Increased R&D Investment, Decreased Operating Revenue | Establish a special R&D budget to safeguard the baseline for energy consumption optimization per token; and implement multi-region computing power off-peak scheduling to reduce costs and consumption by utilizing regional differences in energy. |
| If any environmental non-compliance incidents occur, it will damage our reputation, resulting in the loss of high-quality customers and hindering investment and financing. . . | Medium | Medium-term | Decreased Operating Revenue | Standardize climate-related environmental information disclosure to enhance the transparency of our low-carbon operations. Establish a normalized public opinion monitoring mechanism for climate issues. |

***Greenhouse Gas Emissions***

The greenhouse gas emission sources during our operations are primarily divided into two categories: (i) purchased electricity; and (ii) indirect greenhouse gas emissions generated by employees' business travel. During our daily operations, in order to reduce greenhouse gas emissions, we implement refined energy-use management and control measures: fully utilizing natural lighting to optimize indoor lighting energy consumption and strictly controlling air conditioning temperature-regulation energy loss. During the Track Record Period, our greenhouse gas emissions are as follows:

| Indicator | Unit | 2023 | 2024 | 2025 |
|---|---|---|---|---|
| Scope 1 greenhouse gas emissions | kgCO2e | 0.00 | 0.00 | 0.00 |
| Scope 2 greenhouse gas emissions—purchased electricity | kgCO2e | 691.69 | 5,612.33 | 18,059.44 |
| Scope 3 greenhouse gas emissions—employee business travel | kgCO2e | — | — | 7,508.01 |
| Total greenhouse gas emissions | kgCO2e | 691.69 | 5,612.33 | 25,567.45 |
| Greenhouse gas emissions intensity | kgCO2e/RMB'000 of revenue | 115.28 | 0.76 | 0.46 |

(1) The calculation of the above indicators is based on relevant regulations such as the GHG Protocol and the IPCC Guidelines for National Greenhouse Gas Inventories. As we are a non-manufacturing enterprise and considering our actual operations, Scope 1 direct greenhouse gas emissions are not applicable as there is no direct energy consumption involved.

(2) Scope 2 indirect emissions primarily consist of greenhouse gas emissions generated from our purchased electricity. The calculation of relevant indicators refers to the average emission factors of China's provincial power grids.

(3) Scope 3 greenhouse gas emissions — Category 6 Business Travel covers employees' business travel via aviation and high-speed rail. Due to current data collection constraints, only the employee travel data for the year 2025 has been fully collected at this stage. Going forward, we will improve the travel ledger management to continuously strengthen the data foundation for carbon emission accounting.

## Social

### *Labor Relations*

We strictly comply with laws and regulations such as the Labor Law of the People's Republic of China (《中華人民共和國勞動法》), and have established internal policies to protect employees' legitimate rights and interests, fostering harmonious, stable, and mutually beneficial labor relations. We strictly prohibit the employment of child labor and forced labor, and have established a full-process employment compliance review mechanism. Furthermore, we eliminate any differential treatment of job applicants and current employees based on factors such as gender, age, or ethnicity, adhering to the principles of equal opportunity and merit-based recruitment. As of December 31, 2025, we achieved a 100% labor contract signing rate and 100% social insurance coverage. We employed 1 disabled employee and provided matching job opportunities tailored to their abilities and needs. The detailed breakdown of our workforce structure is set out below:

| | | As of December 31 | | |
|---|---|---|---|---|
| Category | Subgroup | 2023 | 2024 | 2025 |
| Gender | Male employees | 16 | 32 | 74 |
| | Female employees | 1 | 11 | 26 |
| Age | ≤ 30 years (incl.) | 11 | 17 | 32 |
| | 31—49 years | 6 | 23 | 66 |
| | ≥ 50 years (incl.) | 0 | 3 | 2 |

We organize all-staff meetings on an irregular basis each year to collect and communicate employee grievances and feedback, continuously enhancing employees' sense of belonging and workplace well-being. In addition, we formulate competitive and equitable remuneration and benefits packages by integrating factors such as market benchmarks, employee performance, and job contributions, which encompass basic salary, variable bonuses, social security and housing provident funds, annual bonuses, long-term incentives, and fringe benefits. Staff turnover stood at 0%, 15.69% and 7.34% in 2023, 2024 and 2025 respectively, mainly due to personal career development reasons.

***Development and Training***

We adhere to the people-oriented and practice-driven talent cultivation philosophy, attaching great importance to employee development. We implement tiered and categorized cultivation closely aligned with our strategic development, relying on mentorship and hands-on project experience to train and develop our workforce. We continuously enhance our learning and empowerment system, with a training framework covering management, general, and professional competencies. Our training methods are primarily internal, supplemented by external training, and the training content includes specialized business and technical training, as well as safety training. During the Track Record Period, the employee training coverage reached 100% annually.

***Occupational Health and Safety***

We attach great importance to the occupational health and safety of our employees, providing them with a safe, healthy, and comfortable working environment. We have established an employee mental health care mechanism, with designated personnel conducting communications and emotional guidance. To further strengthen our emergency response capabilities for sudden health events, we have purchased AED emergency equipment and placed it in a core area of the office, and have invited external trainers to conduct specialized training on CPR and AED utilization for all employees. As of the Latest Practicable Date, we have obtained the ISO 45001:2018 Occupational Health and Safety Management System certification. During the Track Record Period, we experienced no major safety incidents, nor were there any work-related fatalities or workplace injuries.

***Supply Chain Management***

We are committed to building a stable, compliant, and efficient supply chain system. We formulate internal management policies, and implement rigorous onboarding and dynamic management mechanisms. When onboarding new suppliers, they must undergo multiple internal control processes, including technical reviews. We evaluate core suppliers including IDC operators, cloud vendors and computing hubs focusing on operational capacity and service sustainability, prioritizing partners with sound ESG governance. The assessment covers qualifications, information security, quality management and environmental management. We also prefer suppliers sourcing renewable power. In our daily supplier management, we have established routine supplier supervision and assessment mechanisms. Meanwhile, the procurement team regularly collaborates with the technical team to evaluate and confirm suppliers' service capabilities on a monthly basis. For non-compliant suppliers, we will hold them accountable according to contract terms. In addition, we require suppliers to sign an Integrity Agreement (《廉潔協議書》) to explicitly prohibit commercial bribery and improper transfer of benefits, and we have established dedicated reporting channels to ensure that the supply chain system operates in a standardized, transparent, and benign manner.

***Product Liability***

We have formulated systems such as the Product Quality Emergency Response Plan (《產品質量應急處置預案》) and the Business Continuity Management Measures (《業務連續性管理辦法》), establishing multi-level emergency response workflows that clearly define the reporting pathways, handling time limits, and division of responsibilities for different levels of incidents to ensure that sudden quality events can be rapidly responded to and effectively resolved. In light of the technical characteristics and compliance requirements of the AI model inference infrastructure industry, we have established a

quality management system and an information security control system to effectively safeguard the legitimate rights and interests of customers, as well as the safe and stable operation of businesses. As of the Latest Practicable Date, we have obtained the ISO 9001:2015 Quality Management System Certification, the Capability Maturity Model Integration Level 3 Certification, the ISO 42001:2023 Artificial Intelligence Management System Certification, and have received 8 other AAA certificates, including the "Quality-Focused and Trustworthy Enterprise." We conduct regular system surveillance audits and internal quality reviews annually to optimize and refine our quality management measures in a timely manner, building a closed-loop quality management system across the entire process.

In addition, we have constructed a tiered and multi-dimensional customer service system and formulated the Customer Rights and Interests Protection Management System (《客戶權益保護管理制度》) to elevate customer satisfaction. As of the Latest Practicable Date, we have not experienced any major incidents arising from product liability, and there have been no related claim incidents causing major personal casualties or property losses.

***Information Security and Data Privacy Protection***

We strictly comply with laws and regulations such as the Cybersecurity Law of the People's Republic of China (《中華人民共和國網絡安全法》), and have formulated the information security management and privacy protection framework. The Platform Engineering Department serves as the centralized governing body for our cybersecurity functions and coordinates the management of information security. In addition, we deploy full-time network management personnel to execute practical tasks such as daily network operation and maintenance and safety implementation. As of the Latest Practicable Date, we have obtained the ISO 27001:2022 Information Security Management System certification, the ISO 27701:2019 Privacy Information Management System certification, and the Multi-Level Protection Scheme for Information System Security filing, building an information security protection system that conforms to national laws, regulations, and industry standards.

We build a closed-loop security control system around the entire data life-cycle, formulating specialized standards such as data classification and grading, as well as permission management. We assign access permissions based on the principle of least privilege and obtain informed consent from users where required. At the technical level, we define resource security boundaries relying on computing and network isolation, adhere to minimized storage requirements, and delete or anonymize customer data after project completion. For public cloud services, we provide high-level security protection. In the private deployment mode integrated hardware and software products are deployed on the client side, and relevant software, hardware, models, and data never leave the customer's internal network, fundamentally eliminating data leakage risks.

***Intellectual Property Protection***

We attach great importance to intellectual property layout and the protection of independent innovation. We have formulated the Patent Management System (《專利管理制度》) to protect intellectual property. We encourage employees to actively apply for intellectual property rights and guide technological innovation through cash bonus incentives. For critical positions, we effectively prevent the risk of losing core technologies and trade secrets through agreed non-compete provisions, confidentiality, and intellectual property transfers.

***Community Investment***

Leveraging our proprietary technologies to advance ecosystem development and deliver social value, we focus on inclusive AI, developer ecosystem collaboration, public-interest research support, and open-source community enablement. We have made certain open-source models under 9B parameters freely available to all platform users, enabling developers, students, independent creators, and small-to-medium teams to access AI models at low cost, thereby lowering the upfront investment and barriers to AI innovation. In addition, we provide free token resources support to our public welfare and scientific research partners subject to pre-approved allocation limits, effectively reducing the cost of using AI in scientific research and public welfare projects. As of the Latest Practicable Date, we have engaged in more than 10 key support projects, covering diverse scenarios such as startup camps, campus sharing sessions, hackathons, and overseas roadshows. With token grants as the core method of support, a cumulative total of approximately RMB1.4 million in token grants has been distributed, significantly lowering the thresholds for innovative practice and entrepreneurial launch.

### *Business Ethics*

We advocate a corporate culture of honesty and integrity, strictly comply with all relevant laws and regulations, and foster an anti-fraud corporate culture and environment. We clearly define professional ethics standards for employees in the Employee Handbook (《員工手冊》) and labor contracts, integrate relevant content into the induction training for new employees, and provide accessible and confidential channels for reporting any improper conduct. During the Track Record Period, the coverage rate of employees and directors receiving professional ethics and anti-corruption training, as well as the coverage rate of suppliers to whom our professional ethics and anti-corruption policies were communicated, both remained at 100%. As of the Latest Practicable Date, we have not received any reports regarding incidents related to business ethics.

## COMPETITION

Explosive growth characterizes China's emerging token supply market as it transitions to the rapid scaling of large-scale AI inference demand, driven by the evolution of AI applications toward continuous task execution and increasingly complex multi-step workflows. The size of China's token supply market, as measured by token throughput, grew by 1,602.6% from 2024 to 2025, and is expected to reach approximately 53.2 quintillion tokens by 2030, representing a CAGR of 638.3% from 2025 to 2030.

We differentiate ourselves as an independent ecosystem token supply platform. Unlike closed ecosystems that restrict users to their proprietary computing resources and models, we maintain a neutral and open position. We connect diverse, heterogeneous computing resources and multi-model ecosystems without directly competing with downstream application vendors. Through our deep technical capabilities in cross-chip adaptation, inference engine optimization, and heterogeneous computing resource orchestration, we maximize token production efficiency and ensure stable, scalable delivery across multiple deployment modes. In 2025, we were the fourth largest token supply platform in China in terms of annual token throughput, with a market share of 1.5% and ranked first among all independent ecosystem token supply platforms in China. For details, see "Industry Overview."

## AWARDS AND RECOGNITION

During the Track Record Period and up to the Latest Practicable Date, we received awards and recognition in respect of our services, technology and innovation, significant ones of which are set forth below:

| Award/Recognition | Award Authority | Award Year |
|---|---|---|
| 2026 Super AI Leader (SAIL) Award Top 30 (2026卓越人工智能引領者獎 Top 30) | World Artificial Intelligence Conference (世界人工智能大會) | 2026 |
| National High-Tech Enterprise (國家高新技術企業) | National Leading Group Office for the Administration of High-Tech Enterprise Certification (全國高新技術企業認定管理工作領導小組辦公室) | 2025 |
| Beijing Digital Infrastructure Technology Benchmark Enterprise (北京市數字基礎技術標桿企業) | Beijing Software and Information Service Industry Association (北京軟件和信息服務業協會) | 2025 |

| Award/Recognition | Award Authority | Award Year |
|---|---|---|
| Beijing Specialized, Refined and Novel SME (北京市專精特新中小企業) | Beijing Municipal Bureau of Economy and Information Technology (北京市經濟和信息化局) | 2025 |
| Beijing Innovative SME (北京市創新型中小企業) | Beijing Municipal Bureau of Economy and Information Technology (北京市經濟和信息化局) | 2025 |
| Beijing Science and Technology-Based SME (北京市科技型中小企業) | Beijing Municipal Science and Technology Commission (北京市科學技術委員會) and Zhongguancun Science Park Management Committee (中關村科技園區管理委員會) | 2025 |
| 2025 Model-as-a-Service (MaaS) Benchmark Case (2025年度模型服務MaaS標桿案例) | China Artificial Intelligence Industry Development Alliance (AIIA) (中國人工智能產業發展聯盟) | 2025 |
| 50 Smartest Companies (50家聰明公司) | MIT Technology Review (《麻省理工科技評論》) | 2025 |
| 21 High-Growth Innovative Companies (21家高成長性創新公司) | China Entrepreneur Magazine (《中國企業家》雜誌) | 2025 |
| China's Most Valuable AGI Innovation Institutions TOP 50 (中國最具價值AGI創新機構TOP 50) | Geek Park (極客公園) | 2024 and 2025 |
| New Sprout 50 & Artificial Intelligence 50 (新芽50 & 人工智能50) | Zero2IPO Holdings Inc. (清科控股) and PEdaily.cn (投資界) | 2025 |
| 2024 Most Noteworthy AIGC Enterprises & AIGC Products Dual Award (2024年中國最值得關注AIGC企業 & AIGC產品雙料獎項) | Qbit (量子位) | 2024 |

## EMPLOYEES

As of December 31, 2025, we had 100 full-time employees, all of whom were located in Chinese mainland. In addition, 41 full-time employees hold a master's degree or above. The following table sets forth the number of our employees by function:

| Employee Function | Number of Employees | % of Total |
|---|---|---|
| Research and Development | 67 | 67.0 |
| Sales and Marketing | 21 | 21.0 |
| Operations | 5 | 5.0 |
| General and Administrative | 7 | 7.0 |
| **Total** | **100** | **100.0** |

Our success deeply rests with our ability to attract, retain and motivate qualified talent, and we believe that our high-quality talent pool is one of our core strengths and competitive advantages. We recruit according to high standards and rigorous procedures, using various methods, including online recruitment, internal referral programs and third-party recruiters, to select suitable candidates for relevant positions based on our talent needs.

We invest in continuing education and training programs, including regular and tailor-made internal and external training, for our employees to improve their professional knowledge and management skills, upgrade their skill sets and keep abreast of the industry standards in their respective positions. We also organize activities to help our employees gain a better understanding of our culture.

In line with PRC laws, we participate in government-mandated employee benefit plans, including social insurance for pensions, medical care, unemployment, work-related injury, maternity, and housing funds.

We believe we have a positive working relationship with our employees. None of our employees are represented by labor unions. Throughout the Track Record Period and up to the Latest Practicable Date, we experienced no strikes, work stoppages, or labor disputes that materially affected our business operations.

## INSURANCE

As of the Latest Practicable Date, we maintained commercial insurance coverage for all full-time employees who have completed their probationary period. The commercial insurance mainly includes supplemental medical insurance, and other covered matters include accidents, traffic, etc. We do not otherwise maintain any business interruption insurance or product liability insurance, which are not mandatory under PRC laws and in line with general market practices. We do not maintain keyman life insurance, insurance policies covering damages to our network infrastructures or information technology systems or any insurance policies for our properties.

During the Track Record Period, we did not make any material insurance claims in relation to our business. Any uninsured occurrence of business disruption, litigation or natural disaster, or significant damages to our uninsured equipment or facilities could have a material adverse effect on our results of operations. For details, see "Risk Factors — Risks Related to Our Business and Industry — Our limited insurance coverage could expose us to significant costs and business disruption."

## LICENSE, APPROVALS AND PERMITS

We maintain required licenses, permits and approvals and monitor ongoing compliance to ensure all necessary authorizations are in place. During the Track Record Period and up to the Latest Practicable Date, no material unexpected or adverse changes that could adversely affect the maintenance and renewal of our material licenses, permits, approvals and certificates had occurred since the dates of issue of the relevant regulatory approvals for our business operation. The following table sets forth the details of the material licenses and permits necessary for our business operations:

| License/Permit | Entity Holding the License/Permit | Grant Date | Expiration Date |
|---|---|---|---|
| Value-Added Telecommunications Business License (增值電信業務經營許可證) | The Company | August 2024 | July 2029 |
| Domestic Deep Synthesis Service Algorithm Filing (境內深度合成服務算法備案) | The Company | September 2025 | N/A |

| License/Permit | Entity Holding the License/Permit | Grant Date | Expiration Date |
|---|---|---|---|
| Domestic Deep Synthesis Service Algorithm Filing (境內深度合成服務算法備案) | The Company | March 2026 | N/A |
| Generative AI Service Registration (生成式人工智能服務登記) | The Company | December 2025 | N/A |
| Information System Security Classified Protection Filing Certificate (信息系統安全等級保護備案證明) | The Company | October 2024 | N/A |

## PROPERTIES

As of the Latest Practicable Date, we did not own any properties in the PRC. As of the Latest Practicable Date, we primarily leased three properties in the PRC that are in actual use, with an aggregate gross floor area of 2,956.74 sq.m. as our offices. We believe that our current facilities are adequate to meet our current needs.

## LEGAL PROCEEDINGS AND COMPLIANCE

During the Track Record Period and up to the Latest Practicable Date, we had not been involved in any actual or pending legal, arbitration or administrative proceedings (including any bankruptcy or receivership proceedings) that we believe would have a material adverse effect on our business, results of operations, including the R&D of our Specialist Technology Products, financial condition or reputation and compliance.

During the Track Record Period and up to the Latest Practicable Date, we did not commit any material non-compliance of the laws and regulations, or experience any systemic non-compliance incident which, taken as a whole, in the opinion of our Directors, is likely to have a material adverse effect on our business, results of operations and financial condition. During the Track Record Period and up to the Latest Practicable Date, we had complied with the relevant PRC laws and regulations in all material respects.

## RISK MANAGEMENT AND INTERNAL CONTROL

We have established and currently maintain risk management and internal control systems consisting of policies and procedures that we consider to be appropriate for our business operations. We are dedicated to continuously improving these systems. We have adopted and implemented risk management policies in various aspects of our business operations. Our Board of Directors is responsible for the establishment and updating of our internal control systems, while our senior management monitors the daily implementation of the internal control procedures and measures with respect to each subsidiary and functional department. Our Directors are of the view that we have adequate and effective internal control procedures to ensure compliance with relevant laws and regulations going forward.

### Human Resource Risk Management

We have established internal control policies that cover all aspects of human resource management, including recruitment, training, professional ethics, and legal compliance. Our industry is in dire need of experienced employees, especially R&D personnel. The departure of key personnel may have an adverse effect on us.

All our employees have entered into employment agreements with us containing confidentiality and non-competition clauses. We also require employees to adhere to higher professional ethical standards. We provide all employees with an employee handbook which includes a code of conduct that each employee must adhere to.

### Financial Reporting Risk Management

We have implemented a series of accounting policies for financial reporting risk management and have established strict internal reimbursement, financial reporting management and approval policies. Specifically, the finance department implements specific review and verification procedures for invoices, drafts, bills, and other financial documents to ensure the authenticity of the original documents we receive and use. The finance department also checks whether the amounts and times shown on the documents are consistent with the relevant contracts. We have a strict internal approval process, and almost all approval matters are completed online through the company's internal platform, thereby enabling effective end-to-end monitoring while operating efficiently.

### Legal Compliance and Intellectual Property Risk Management

Our operation risk management involves compliance with the PRC laws and regulations, especially laws and regulations relating to the information service industry, as well as protecting intellectual property and avoiding liability for infringing third-party intellectual property rights. We have a team of experienced legal professionals to ensure our compliance and control intellectual property-related risks. Our legal team is responsible for approving contracts, monitoring updates in the PRC laws and regulations, and ensuring that business operations continue to comply with the relevant laws and regulations. Our legal department also assists in ensuring the timely filing of all necessary trademark, copyright, and patent applications with the relevant authorities.