GLOSSARY OF TECHNICAL TERMS

This glossary contains definitions of certain technical terms used in this document in connection with us and our business. These may not correspond to standard industry definitions and may not be comparable to similar terms adopted by other companies.

"AGI"	artificial general intelligence, a theoretical form of artificial intelligence that possesses the ability to understand, learn, and apply knowledge across a wide range of tasks at a level equal to or beyond human capabilities
"AI"	Artificial Intelligence, the simulation of human intelligence processes by machines, especially computer systems
"AI agent" or "agent"	an artificial intelligence system capable of autonomous understanding, planning, and executing complex workflows or tasks, transitioning from simple text generation to autonomous execution
"AI application"	software applications that utilize artificial intelligence technologies to perform specific tasks or provide services
"AI capability"	the sense of intelligence or smart capabilities delivered by artificial intelligence systems
"AI infrastructure"	the underlying hardware and software resources, including computing power, storage, and networking, required to develop, train, and deploy AI models
"AI model"	a mathematical or computational model trained on data to recognize patterns and make predictions or decisions
"AI-native"	refers to products, services, or infrastructure built from the ground up with artificial intelligence as a core component rather than an add-on feature
"AIGC"	artificial intelligence generated content, the use of AI algorithms to generate new content such as text, images, audio, or code
"API"	application programming interface, a set of rules and protocols that allows different software applications to communicate and interact with each other
"API call"	the process of a client application submitting a request to an application programming interface and receiving a response, typically used by customers to access and utilize AI model inference services on our platform
"ASIC"	Application-Specific Integrated Circuit, an integrated circuit chip customized for a particular use rather than intended for general-purpose use
"CAGR"	compound annual growth rate

"closed ecosystem"	a closed ecosystem AI token supply model typically developed by cloud service providers, large technology companies or model companies based on their own cloud resources, AI models, computing resources and applications. Under this model, token supply services, model invocation, computing resource allocation and application integration are mainly completed within the provider's own ecosystem, forming an internal closed loop
"computational graph"	a directed acyclic graph representing mathematical operations and data flow in a machine learning model
"computing resource"	the processing power and memory provided by hardware to perform computational tasks
"computing resource orchestration"	the automated arrangement, coordination, and management of various computing resources to deliver integrated services
"context"	the background information and prior conversation history provided to an AI model to generate relevant and coherent responses
"context length"	the maximum number of tokens, including both the input prompt and the generated output, that a large language model can process and retain in its memory during a single interaction
"cross-chip"	referring to the involvement or integration of multiple types of chips within a system or architecture
"cross-chip adaptation"	the process of modifying or configuring software and models to run efficiently on different types or multiple chips
"decode"	the stage in AI model inference where the model produces output tokens one by one based on the prefilled context
"dedicated instance"	a cloud computing service offering where specific computing capacity and hardware are reserved for a customer's sole use and are not shared with other users to ensure data isolation and stable performance
"dense model"	a traditional neural network architecture where all parameters and computational nodes are activated for every token or input processed, as opposed to sparse models
"deployment"	the process of installing, configuring, and enabling software or models to be used in a live environment
"domestic chip"	semiconductor chips designed in China
"Dual Carbon"	refers to China's national climate goals of reaching peak carbon emissions before 2030 and achieving carbon neutrality before 2060
"dynamic quantization"	a technique that dynamically reduces the numerical precision of a neural network's weights and activations during inference to lower computing and memory requirements without significant loss of accuracy

GLOSSARY OF TECHNICAL TERMS

"elastic scaling"	the ability of a system to automatically add or remove computing resources in response to changing workloads
"elastic service"	a computing service that can automatically scale its resources up or down based on current demand
"enterprise-grade"	refers to products or services designed to meet the rigorous demands of large organizations in terms of scalability, reliability, and security
"expert parallelism" or "EP"	a distributed computing technique used in AI models where different specialized sub-networks or experts are distributed across multiple computing devices to improve computational efficiency
"GPU"	Graphics Processing Unit, a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images, widely used to handle the massive parallel computational workload of AI model training and inference
"hardware architecture"	the logical structure and physical design of computer hardware components and their interconnections
"hardware utilization"	the percentage of time or capacity that hardware resources are actively used for productive work
"heterogeneous"	consisting of dissimilar or diverse elements; in computing, referring to systems using different types of processors or architectures
"heterogeneous computing"	the use of a variety of different types of processors or cores, such as GPUs, NPUs, and ASICs from diverse manufacturers, within a single computing environment to optimize performance and cost
"heterogeneous computing resource"	computing resources derived from different types of hardware architectures or processors
"high-speed interconnect"	technology or infrastructure that enables fast data transfer and communication between different hardware components or nodes
"IDC"	Internet Data Center, a physical facility used to house computer systems and associated components, such as telecommunications and storage systems
"independent ecosystem"	an independent ecosystem AI token supply platform model that is not bound to a single cloud provider, computing resource, AI model or application. It neutrally connects heterogeneous computing resources, diverse AI models and deployment environments, and provides token supply services across different ecosystems with horizontal adaptation capabilities
"inference"	the process of using a trained AI model to make predictions, solve tasks, or generate outputs based on new user prompts

"inference engine"	a software system designed to execute AI models and generate outputs (inferences) based on input data
"inference optimization"	techniques and processes used to improve the speed and efficiency of model inference while maintaining accuracy
"KV cache"	key-value cache, a memory optimization technique used during the inference of AI models that stores previously computed attention keys and values to prevent redundant calculations in multi-turn interactions
"large-scale deployment"	the installation and operation of systems or models across a massive number of nodes or environments
"latency"	the time delay between a user submitting a request to an AI model and the system providing the corresponding response
"liquid cooling"	a cooling technology that uses liquid coolants to remove heat from computing hardware, improving energy efficiency and density
"load balancing"	the process of distributing workloads across multiple computing resources to ensure optimal resource utilization and prevent overload
"MaaS" or "Model-as-a-Service"	a cloud computing service model that provides users with access to pre-trained AI models via APIs or platforms
"mixture-of-experts" or "MoE"	a deep learning architecture that uses multiple specialized sub-networks known as experts to solve a problem, activating only a subset of experts for any given input to improve computational efficiency
"model capability"	the specific skills and functions that an AI model can perform, such as natural language understanding, image recognition, or reasoning
"model deployment"	the process of integrating a trained AI model into an existing production environment to make predictions or decisions
"model weight"	the learned parameters within an AI model that determine how input data is transformed into output predictions
"multi-model"	referring to the integration or utilization of multiple types of AI models
"multimodal"	referring to AI models or systems capable of processing, understanding, and generating multiple different types of data inputs and outputs, such as text, images, audio, and video
"NPU"	Neural Processing Unit, a specialized microprocessor designed specifically for the acceleration of machine learning algorithms and neural network operations
"observability"	the ability to measure and understand the internal states of a system based on its external outputs, such as logs, metrics, and traces

"on-premise deployment"	a software delivery model where the platform and AI models are installed and operated entirely within the customer's private data center or local infrastructure
"operator"	a fundamental mathematical function or computational routine within a neural network algorithm, the deep optimization of which is critical for adapting AI models to run efficiently on specific chip architectures
"orchestration"	the automated configuration, coordination, and management of complex computer systems, middleware, and services
"pipeline parallelism"	a distributed computing technique that divides an AI model into sequential layers across multiple processors, allowing different micro-batches of data to be processed simultaneously
"PoC" or "proof-of-concept"	a realization of a certain method or idea to demonstrate its feasibility, often used to verify that a concept or theory has the potential for real-world application
"prefill"	the stage in AI model inference where the model processes the input prompt and pre-fills the KV cache to produce the first token
"prefill-decode separation" or "PD separation"	an inference optimization technology that separates the initial processing of a user prompt (prefill) from the subsequent production of output tokens (decode) across different computing nodes to maximize cluster-level scheduling efficiency
"production-grade"	refers to systems or software that are robust, reliable, and secure enough to be deployed in a live, end-user environment
"prompt"	the input text, instruction, or data provided by a user to an AI model to guide its generation of a response or execution of a task
"public cloud-based service"	cloud computing services provided by us over the internet to multiple users
"resource scheduling"	the policy and mechanism of assigning available resources to tasks over time to optimize overall system performance
"serverless"	a cloud computing execution model where the cloud provider dynamically allocates machine resources on demand, allowing customers to build and run AI applications without managing the underlying infrastructure
"SLA"	Service Level Agreement, a commitment between a service provider and a client specifying the expected level of service, typically including metrics such as availability, uptime, and performance stability
"system software"	a type of computer program that is designed to run a computer's hardware and to provide and maintain a platform for applications to run on

"throughput"	the amount of data or number of tokens processed and produced by a computing system or AI model in a given amount of time
"token"	a fundamental unit of data processed by an AI model during inference, analogous to a word or sub-word in natural language, serving as both the generated output and the basic metric for measuring computing consumption and billing
"token consumption"	the number of tokens processed or produced by an AI model during a specific operation or session
"token supply platform"	a system software platform that connects hardware resources, model capabilities, and application demands by organizing, orchestrating, and optimizing the entire token production and delivery process to provide stable, efficient, and scalable token services
"tokens/s"	tokens per second, a standard metric used to measure the production speed and throughput of AI model inference
"training"	the process of using datasets to teach an AI model to recognize patterns and make accurate predictions