This glossary of technical terms contains explanations of certain technical terms used in this Document. As such, these terms and their meanings may not correspond to standard industry meanings or usage of these terms.

"AI" artificial intelligence, a branch of computer science that develops systems capable of performing tasks that typically require human intelligence, such as perception, learning, reasoning, and decision-making

"AIGC" Artificial Intelligence Generated Content, content such as text, visual, and audio created automatically by artificial intelligence technologies without direct human creation

"AI agent" an intelligent system capable of autonomously performing specific tasks on behalf of a user to achieve a proposed goal

"AI Infrastructure" the integrated set of hardware, software, data systems, or cloud resources necessary for training and inferencing AI models

"AI music synthesis model" an artificial intelligence system designed to generate or compose music by learning patterns from existing musical data, enabling creation of original melodies, harmonies, and rhythms

"AI-native" refers to products, services, or systems that are built from the ground up with artificial intelligence as a core component, rather than integrating AI as an add-on. AI-native solutions are designed to leverage AI technologies in their fundamental architecture and functionality

"AI-powered Multi-modal Entertainment Platform" a platform that uses artificial intelligence to enable agents to process and respond to multiple types of inputs, allowing more natural and versatile interactions

"API" a set of rules and tools that allows different software systems to communicate and interact with each other

"API Platform" a system that provides standardized interfaces allowing developers to access and integrate a company's services or data into their own applications

GLOSSARY OF TECHNICAL TERMS

"app" or "application" — application software designed to run on smartphones and other mobile devices

"Asia-Pacific" — A geographic region comprising countries in East Asia, South Asia, Southeast Asia, and Oceania, often including Australia and New Zealand

"attention" — the sophisticated mechanism that enables the foundation model to weigh the importance of different parts of its input data when processing information

"Audio Generation Tool" — software that uses AI to create synthetic audio content, including speech, music, or sound effects, based on user input or predefined parameters

"auto-regressive model" — a type of statistical model that uses past values of a time series to predict future values of that same time series

"B2B" or "2B" or "ToB" — "business-to-business", refers to commercial transactions or relationships between businesses rather than between a business and individual consumers

"B2C" or "2C" or "ToC" — "business-to-consumer", refers to commercial transactions or relationships between a business and individual consumers

"benchmark" — standardized evaluation frameworks used to measure and compare the performance of language models

"CAGR" — compound annual growth rate

"Chat Completions API" — an application programming interface that allows developers to build interactive chat experiences by generating AI-driven responses in a conversational format, typically based on large language models

"chain-of-thought" or "CoT" — a reasoning technique where the model generates intermediate reasoning steps or explanations to solve complex problems

"Diffusion Model" — a type of generative model in machine learning that excels at creating high-quality data, particularly images and text, by gradually adding noise to a data point and then learning to reverse this process

"DiT" diffusion models with transformers, a type of diffusion model that utilizes the transformer architecture as its backbone for image generation

"ELO" a rating system used to assess the relative skill levels of players in competitive games or activities, where a higher score indicates stronger performance

"ESG" environmental, social and governance

"fine-tuning" the process of adapting a pre-trained foundation model to perform a specific task or specialize in a particular domain with higher accuracy and relevance

"Flow-VAE" a hybrid AI model combining Variational Autoencoders and normalizing flows to improve the quality and flexibility of generated data by capturing complex data distributions more effectively

"foundation model" a large-scale, pre-trained model developed on broad and diverse datasets designed to serve as a general-purpose model that can be used for solving a wide variety of tasks

"fps" frames per second, a measure of how many individual frames (images) are displayed or generated each second in a video or animation. Higher fps results in smoother motion

"freemium" a business model that offers basic services or products free of charge while charging for premium features, advanced functionality, or enhanced experiences.

"GDP" gross domestic product, the total monetary value of all goods and services produced within a country's borders during a specific period, commonly used to measure the size and health of a country's economy

"generative AI" a type of artificial intelligence that creates new content by learning patterns from existing data and generating original outputs

"GPU" Graphics Processing Unit, a processor that handles many tasks at once, widely used in AI to speed up model training and data processing

"High and New Technology Enterprise"

refers to an enterprise established in the PRC that is recognized by the competent government authorities as meeting the prescribed criteria in terms of core independent intellectual property rights, research and development capability, technology and product offerings, and revenue composition from high and new technology-related businesses. Enterprises with such designation are entitled to a preferential PRC corporate income tax rate of 15%, subject to fulfilment of the relevant requirements

"HTML"

HyperText Markup Language, the standard language used to create and structure content on the web, defining elements such as text, images, links, and layout in web pages

"IDC"

International Data Corporation (IDC), a global market intelligence, data, and events provider for the information technology, telecommunications, and consumer technology markets

"Image Generation & Music Generation API"

an application programming interface that enables developers to create images and music programmatically using AI models, allowing automated generation of visual and audio content based on user inputs

"image-to-video" or "I2V"

a technology or model that generates video sequences from a single image or a series of images, creating motion and transitions to produce dynamic video content

"inference activities"

the computational processes through which a trained foundation model is deployed to generate outputs or responses based on new user inputs or data. Inference takes place after the model has been trained and involves applying the model's learned parameters to perform reasoning, prediction or content generation in real time. For example, when a user inputs a prompt or message in our MiniMax app and receives a generated text produced by the underlying foundation model, such model computation constitutes an inference activity

"inference cost"

the cost of computational resources needed to use a trained AI model to process inputs and generate outputs "inference latency"

the time delay between providing inputs to an AI model and receiving outputs from the model

"Intelligent Agent Application"

a software application powered by AI that can understand and respond to user inputs in natural language, enabling interactive and human-like conversations for customer service, information retrieval, or other purposes

"large model", "large language model" or "LLM"

advanced AI models trained on massive amounts of text data to understand, generate, and interact using human language. They are capable of performing a wide range of natural language processing tasks, such as text generation, translation, summarization, and question answering

"large-scale hybrid-attention reasoning model"

an AI technique that combines two different attention mechanisms: one is the computationally expensive but high-precision traditional attention (Softmax Attention), and the other is the fast, less resource-intensive linear attention (Lightning Attention). The model uses the fast linear attention for most of the text processing and only activates the high-precision traditional attention for critical parts. This design allows the AI to efficiently process extremely long texts at a lower computational cost, achieving a balance between performance and efficiency

"linear attention" or "linear attention mechanism"

an efficient attention mechanism that reduces the computational complexity of traditional attention from quadratic to linear, enabling faster processing of long input sequences while preserving key information

"long context processing capacity"

the ability of an AI model to understand, retain, and make use of extremely long sequences of input data, such as lengthy texts, conversations, or documents, allowing it to maintain context and coherence over extended interactions with users

"MAU" monthly active user, the number of unique devices that performed at least one action on our AI-native applications and the number of registered user accounts that logged into our web platforms at least once during a given month, including both paying and non-paying users; MAU figures for a given period represent the average monthly active users for the relevant period, calculated as the average of the MAUs for each month within that period

"model-as-a-service" or "MaaS" a cloud-based offering that allows users to access and deploy models via APIs, enabling them to integrate AI capabilities into applications without managing model development, training, or infrastructure

"MCP" Model Context Protocol, a technical standard or framework that defines how context information is structured, exchanged, and managed within AI models, enabling them to better understand and respond based on the surrounding data or dialogue history

"MCP API" an application programming interface that supports the MCP, enabling AI models or applications to exchange, manage, and utilize contextual information efficiently for improved understanding and response generation

"MFU" Model Flop Utilization, a metric of computing power used by a model during training or inference, indicating how efficiently the model utilizes available resources

"model call" the process of sending an input, such as a query, prompt, or data, to a foundation model to obtain an output

"model for speech and music generation" AI model that generates synthetic speech or music based on input data such as text, audio prompts, or musical notation. Speech generation models convert text into natural-sounding spoken language, while music generation models create original musical compositions or continuations in various styles and formats

“MoE” or “Mixture-of-Experts architecture”

Mixture-of-Experts, an AI model architecture that uses a group of sub-networks, or “experts,” where only a subset is activated for each input. This allows the model to scale efficiently by allocating computational resources based on the nature of the task, improving performance while maintaining efficiency

“Multi-Modal Model Suite”

a collection of AI models designed to process and understand multiple types of data inputs, such as text, audio, and video, in an integrated manner, enabling more comprehensive and human-like perception and interaction

“Open Platform”

a publicly accessible interface that allows third-party developers to integrate with a company’s systems or services by providing standardized programming access, enabling the development of applications or services based on the platform’s capabilities

“open-source”

the practice of making a software’s source code publicly available, allowing anyone to view, use, modify, and distribute it, typically under an open-source license

“p”

pixel, a unit that represents a single point in a digital image or display. It is commonly used to measure screen resolution or image dimensions

“parameters”

internal numerical variables or configurations that are learned and adjusted by the foundation model during its training process. These parameters effectively encapsulate the knowledge, patterns, and relationships extracted by the model from the extensive datasets it has been trained on

“Q&A”

question and answer

“R&D”

research and development

“reasoning model”

an AI model designed to simulate logical thinking by drawing inferences, making decisions, or solving problems based on input data, often across multiple steps or contexts "reinforcement learning from human feedback" or "RLHF" — the process that reinforces the model with human demonstrations and preference data to make the model follow instructions and generate detailed, relevant responses

"Scaling Law" — the empirical relationship describing that the performance of models systematically improves as a function of increasing model size, dataset size, and computational resources

"SDK" — Software Development Kit, a collection of tools, libraries, documentation, and code samples that developers use to build applications for a specific platform, system, or service

"semantic space" — a way to represent the meaning of words or concepts as points in a multi-dimensional space, where the dimensions represent semantic features or contexts. It's a computational approach used in natural language processing and related fields to model relationships between words and concepts based on their meaning and context

"SLA" — Service Level Agreements, formal contracts between a service provider and a customer that define the expected level of service, including performance metrics such as uptime, response time, and support quality, as well as remedies if standards are not met

"softmax attention" — an algorithm that helps AI models focus on the most relevant parts of the input by assigning weights using the softmax function

"text-to-speech" or "TTS" — A technology that converts written text into spoken audio, allowing machines to read text aloud in a natural and intelligible voice

"test-time compute" — refers to the amount of computational power used by an AI model when it is generating a response or performing a task after it has been trained

"token" — a unit of text as the fundamental element for input processing and output generation for models

"Transformer" the sophisticated neural network architecture that can efficiently processes sequential data, such as text, by utilizing attention mechanisms to prioritize key input elements

"URL" Uniform Resource Locator, the address used to locate and access resources, such as web pages, on the internet

"Video Generation API" an application programming interface that allows developers to programmatically create or control the generation of video content using AI models, typically from inputs like text, or audio

"video generation model" AI model designed to create synthetic video content by generating sequences of images over time, often based on text descriptions, images, or other input data. This model learns motion, visual consistency, and temporal coherence to produce realistic or stylized video outputs