Date: December 31, 2025

For and on behalf of

China Insights Consultancy

Vincent Chen

Name: Vincent Chen

Title: Executive Director

灼识咨询
China Insights Consultancy

Industry Report on the Global Foundation Model Market

China Insight Consultancy introduction, methodologies and assumptions

China Insights Consultancy was commissioned to conduct research, provide an analysis of, and to produce a report on the global foundation model industry, and other related economic data, at a fee of USD 115,000. The commissioned report has been prepared by China Insights Consultancy independent of the influence of The Company and other interested parties.

China Insights Consultancy is an investment consulting company originally established in Hong Kong. Its services include industry consulting services, commercial due diligence, strategic consulting, and so on. Its consultant team has been tracking the latest market trends in e-commerce, TMT, marketing and advertising, culture and entertainment, cloud communication, IT, chemicals, consumer goods, agriculture, and industry, finance and services, healthcare, transportation, etc., and possesses the most relevant and insightful market intelligence regarding these industries.

China Insights Consultancy undertook both primary and secondary research using a variety of resources. Primary research involved interviewing key industry experts and leading industry participants. Secondary research involved analyzing data from various publicly available data sources, annual reports published by relevant industry participants, industry associations, China Insights Consultancy's own internal database, etc.

The market projections in the commissioned report are based on the following key assumptions: (i) the overall global social, economic, and political environment is expected to maintain a stable trend during the forecast period; (ii) certain key industry drivers are expected to continue driving market growth during the forecast period; and (iii) there is no extreme force majeure or unforeseen industry regulations in which the market may be affected either dramatically or fundamentally during the forecast period.

All statistics are reliable and based on information available as of the date of this report. Other sources of information, including those from the government, industry associations, or market participants, may have provided some of the information on which the analysis or its data is based.

All the information about MiniMax is sourced from MiniMax's own audited report or management interviews. China Insights Consultancy is not responsible for verifying the information obtained from MiniMax.

Terms and abbreviations

Agent: 代理智能体

AGI: Artificial General Intelligence 通用人工智能

AI: Artificial Intelligence 人工智能

API: Application Programming Interface 应用程序接口

Attention mechanism: 注意力机制

BERT: Bidirectional Encoder Representations from Transformers 双向编码器表示

CAGR: compound annual growth rate 年均复合增长率

CLIP: Contrastive Language–Image Pretraining 对比语言图像预训练

CoT: Chain of Thought 思维链

Context Window: Context Window 上下文窗口

Closed-source Model: 闭源模型

CPC: Cost per click 每次点击成本

CPM: Cost per mille 每千次展示成本

DiT: Diffusion Transformer 扩散变换器

DPO: Direct Preference Optimization 直接偏好优化

ELO: Elo Rating System Elo评分系统

Fine-tuning: 模型微调

FlashAttention: 高效注意力计算算法

Generalization: 泛化能力

Linear Attention: 线性注意力机制

LLM: Large language model 大语言模型

MaaS: Model-as-a-Service 模型即服务

MoE: Mixture of Experts 专家混合架构

Open-source Model: 开源模型

Pre-training: 预训练

Prompt: 提示输入

RLHF: Reinforcement Learning from Human Feedback 基于人类反馈的强化学习

TAM: Total Addressable Market 潜在市场规模

Test-time Compute: Test-time Compute 推理阶段计算

Token: 词元

Transformer: 基于自注意力机制的神经网络架构

ViLBERT: Vision-and-Language BERT 视觉与语言BERT模型

VisualBERT: Visual BERT 视觉BERT模型

Source: China Insights Consultancy

Table of Content

Foundation model is increasingly driving intelligent transformation across key sectors of society and gradually permeating every aspect of people's daily lives.

Real-world adoption across diverse industry scenarios

Application scenarios

Smart home

Transportation

Finance

Healthcare

Specific user cases

Content generation
- Generating text, audio, or visuals automatically to support creative tasks at home.

Voice-controlled assistants
- Smart assistants powered by LMs respond to spoken commands

Home automation systems
- Managing lighting, temperature, and appliances for convenience and energy efficiency.

Autonomous driving systems
- Processing sensor data and support decision-making for self-driving vehicles.

Intelligent traffic management
- Analyzing real-time transportation data to optimize traffic flow and reduce congestion.

Real-time route optimization
- Helping navigation systems adjust routes based on live traffic and road conditions.

Fraud detection models
- Detecting anomalies in transactions and flag suspicious behavior to prevent fraud.

Automated credit scoring
- Evaluating creditworthiness using diverse data sources for faster, data-driven decisions.

Investment advisory
- generate personalized financial recommendations to support better investment choices.

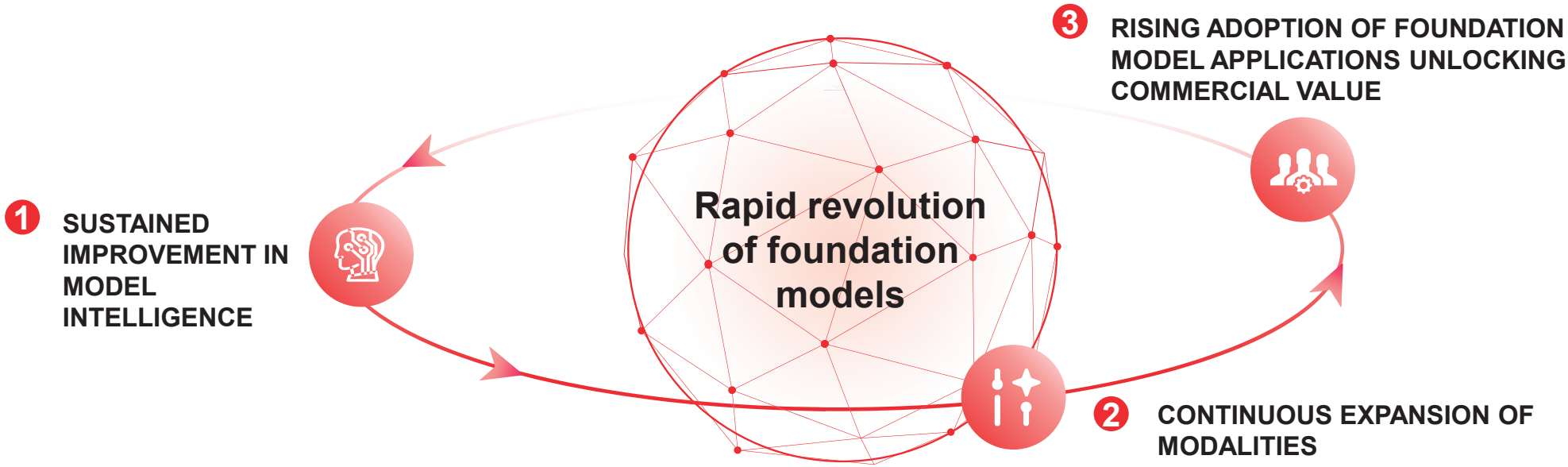
Medical imaging analysis
- Assisting in interpreting medical images and identifying potential abnormalities.

LM-assisted diagnosis
- Supporting clinicians with diagnosis suggestions and relevant medical knowledge.

Robotic-assisted surgeries
- Enhancing surgical precision by assisting with robotic system control and planning.

Source: China Insights Consultancy

Benefitting from significant improvements in model scale and intelligence, the expansion of multi-modal capabilities, and the acceleration of commercialization, the industry has evolved at a remarkable pace.



Artificial intelligence (AI), an intelligent system integrated into human society, is not only unlocking substantial productivity, but also enriching creativity. Today, AI permeates various aspects of both people's professional and personal lives, from social media content recommendations, chatbots, and intelligent personal assistants, to autonomous driving systems, intelligent risk control models, and AI-assisted medical diagnostics. AI has become the core driving force behind the intelligent transformation of society and industries worldwide.

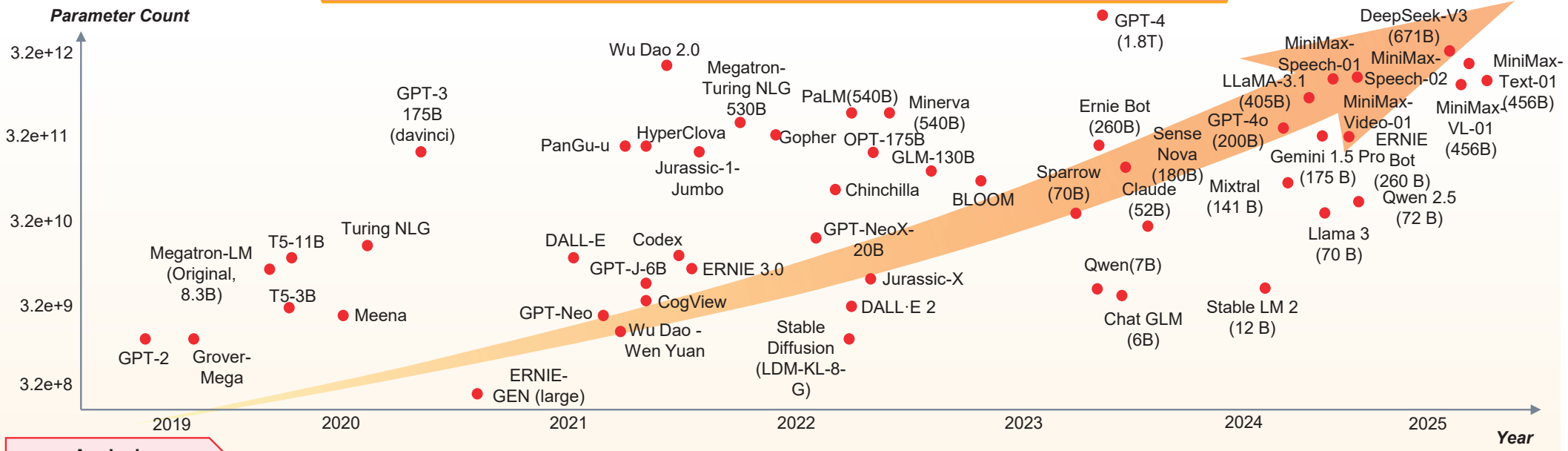
Foundation models in the past three years represent a significant technological paradigm shift compared to previous generations of AI – an inevitable trend fueled by societal developments. Traditional AI centered around small-scale models, which were custom-trained for application-specific scenarios. However, the goal of AGI is to enable intelligence that can perform the full range of human intellectual tasks. This requires AI to be more general-purpose, as user needs are becoming increasingly personalized and diverse. Foundation models are designed to address this challenge by offering strong scalability and generalization capabilities, and represent the most promising path towards achieving AGI.

Over the past three years, the field of foundation models has undergone rapid evolution. Benefitting from significant improvements in model scale and intelligence, the expansion of multi-modal capabilities, and the acceleration of commercialization, the industry has evolved at a remarkable pace.

Source: China Insights Consultancy

Foundation models have scaled up dramatically in recent years in terms of parameters used, with significant performance improvements.

Development and Parameter Scale of Foundation Models



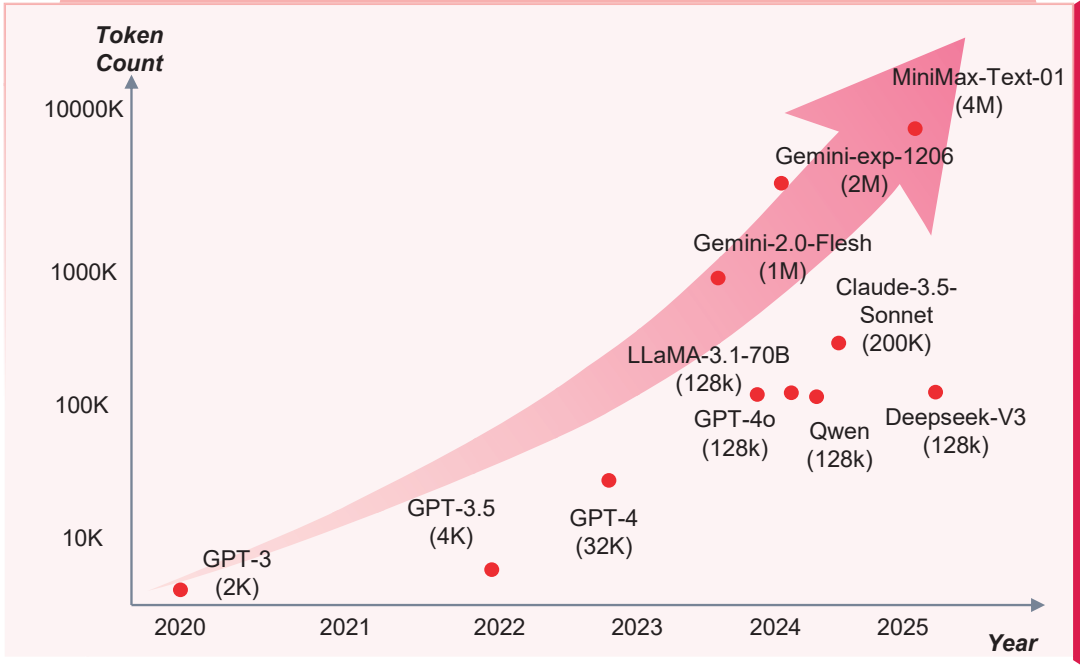
Analysis

- OpenAI's GPT-3, released in 2020, featured 175 billion parameters, while GPT-4 released in 2023 demonstrated capabilities far exceeding its predecessor. GPT-4 scored in the top 10% of test-takers on a simulated bar exam, while GPT-3.5 ranked in the bottom 10%, illustrating that the reasoning and comprehension capabilities of foundation models are approaching or even partially surpassing human levels.
- Since then, more advanced models, such as GPT-4o, Gemini 2.5 Pro, and Claude 3.7, have been released, further pushing the boundaries of both model scale and intelligence. A major breakthrough was the widespread adoption of the Mixture-of-Experts (MoE) architecture, which incorporates multiple expert sub-networks with a gating network that dynamically routes input to the most relevant experts. It is able to enfoundation model scale while maintaining relatively low per-token computational cost and latency.

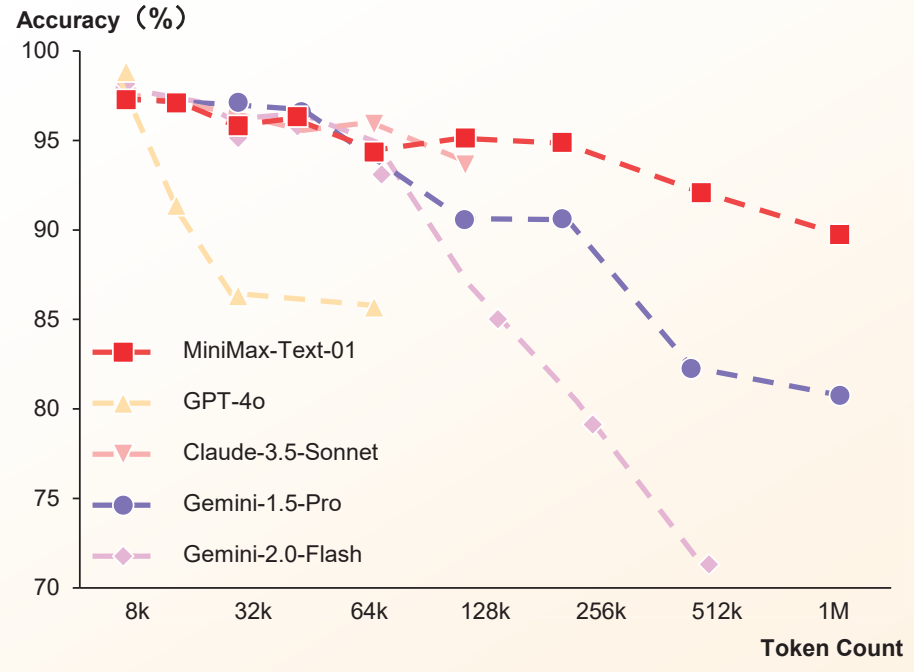
Source：HAI，China Insights Consultancy

New models are not only more intelligent, but also capable of memorizing and processing more content.

Supported Token Count of Context Windows in Text-Based Models



Long-Text Performance of Text-Based Models



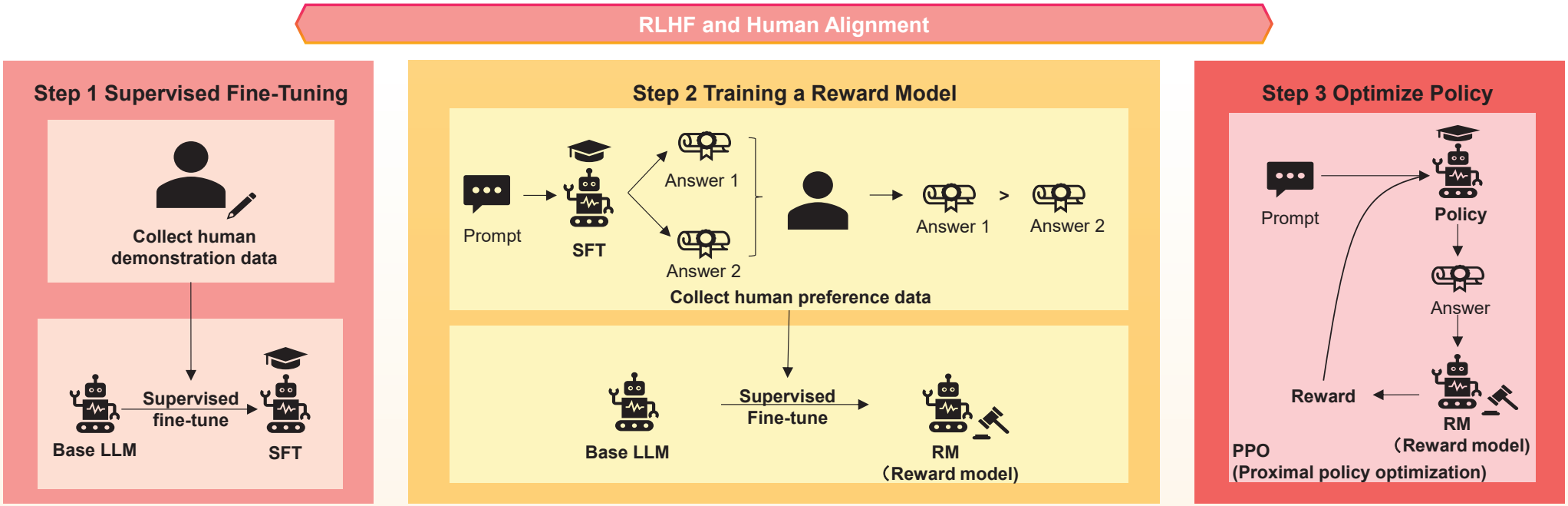
Analysis

- GPT-3 supported a context window of approximately 2,048 tokens; GPT-3.5 increased this to 4,096; GPT-4 offered a context window of up to 32,768 tokens. In 2023, Anthropic's Claude expanded the context window to 100,000 tokens, enabling it to read and answer questions on ultra-long documents.
- However, this also led to significantly higher inference costs. To address this challenge, the industry has implemented efficient underlying model architectures and retrieval-augmented generation techniques. Among the most notable innovations are improvements to the attention mechanism – for example, MiniMax Text-01 pioneered the first large-scale application of the innovative linear attention mechanism

Note: Average accuracy over 13 Ruler task tests.

Source: China Insights Consultancy

RLHF (reinforcement learning from human feedback) allows foundation models to be more receptive to user prompts.



Analysis

- Since the introduction of ChatGPT, RLHF has become a standard within the industry: pre-trained models are reinforced using human feedback and preference data, enabling them to learn how to be more instruction adherent and provide more detailed, useful responses. OpenAI, for example, applied RLHF to refine GPT-3.5 into ChatGPT, which is capable of delivering coherent, tailored, and practical answers.
- This technology has greatly improved model usability in dialogue-based applications and has also inspired alternative alignment solutions such as Anthropic's "Constitutional AI," which guides model behavior through predefined principles rather than detailed human feedback. Foundation models with human alignment have demonstrated remarkable improvements in factual accuracy, tone moderation, and the ability to decline inappropriate requests.

Source: China Insights Consultancy

The CoT prompting technique significantly enhanced foundation models' performance on complex tasks such as mathematical problem-solving, and common-sense reasoning.

Overview of CoT and reasoning models

CoT Action: Convert 30° C to Fahrenheit (° F)?

Model	Process	Result
Simple LLM approach	No immediate steps	The temperature is 86 ° F
CoT Reasoning	Apply Formula ° F = (° C + 9/5) + 32 → First Multiply by 9/5 (30° C × 9/5) = 54 → Add 32 to the multiplied result ° F= 54 + 32 = 85	Temperature in Fahrenheit (° F) is 86 degrees.

Analysis

- The CoT prompting technique, which emerged in 2022, significantly enhanced foundation models' performance on complex tasks such as mathematical problem-solving, and common-sense reasoning, by guiding models to generate intermediate reasoning steps.
- A major shift occurred in late 2024, when certain models began to be specifically designed to break down problems step-by-step during inference. These models allocate greater computational resources in reasoning to perform iterative reasoning, reflection, and output optimization. Reasoning is increasingly regarded as a distinct, computable process that can be optimized and extended, rather than merely an emergent property due to foundation model scale.
- Reasoning models built on test-time compute (such as OpenAI o1and DeepSeek R1) are explicitly trained with additional compute during the reasoning process, often via reinforcement learning, to conduct iterative or structured reasoning during test-time. This trade-off between increased reasoning cost/latency and higher reasoning quality may lead to a divergence in future model development: one class of models will be optimized for fast, factually accurate responses, while the other will focus on deeper, more resource-intensive reasoning, with test-time compute being a key variable.

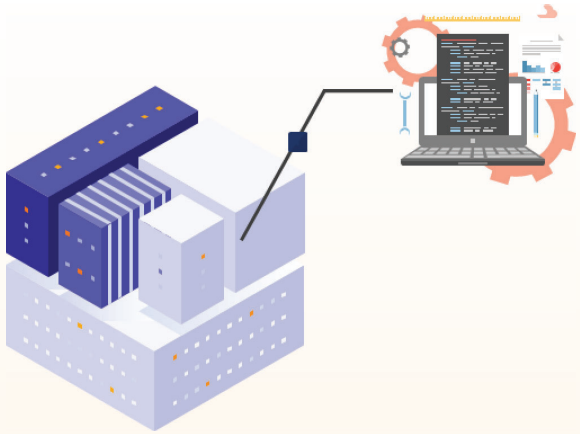
Source: China Insights Consultancy

The current wave of foundation models has ushered in a new paradigm centered around AI agents – models that can plan autonomously and call on external tools to accomplish more complex tasks.

Overview of "AI Agent" and external tool usage

OpenAI

Gemini

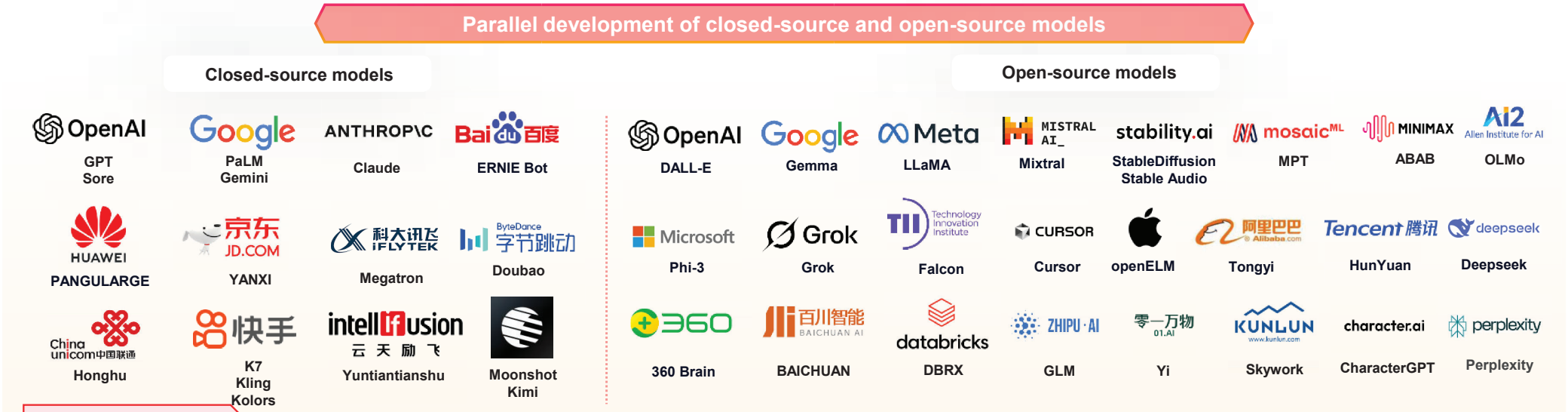


Analysis

- In 2023, OpenAI introduced plug-in and function calling capabilities for GPT-4, allowing the model to invoke external tools such as web browsers, and search engines, overcoming the limitations of models that could previously only operate within the bounds of their training data. For example, when a user enquires about real-time stock market trends, ChatGPT can access real-time data via a browser plug-in before generating a response. Similarly, when asked to solve a complex math problem, the model can call Python to execute code for precise results.
- Google's Gemini took this further by integrating code execution capabilities directly into its API, allowing the model to run code autonomously within a sandbox environment and refine its answers based on the output. Such functionality transforms foundation models into active intelligent agents rather than passive responders. Other tool use includes generating structured outputs for other systems to read, and integration with knowledge retrieval systems.
- In summary, the rise of agentic AI marks a new stage in the evolution of foundation model applications. Rather than operating in isolation, models now function as decision-making hubs, orchestrating various resources to accomplish user tasks. This has not only greatly expanded the scope of AI applications, but also represents a critical step towards AGI.

Source: China Insights Consultancy

Over the past few years, the foundation model industry has seen both closed-source and open-source models developing in tandem.



Analysis

- On the other hand, companies like Meta have promoted open-source models, leading to a boom in the foundation model community in 2023. Meta released LLaMA for academic research in early 2023, followed by the open-sourced update of LLaMA 2 in July, significantly lowering the barrier to entry for derivative model development.
- As a result of these open-source development, the academic community have conducted substantial research, working together with the industry to accelerate progress. Projects like Stanford's Alpaca and LMSYS's Vicuna, both fine-tuned from LLaMA, democratized the exploration of instruction adherence models. At the underlying architecture level, the academic community has also introduced a range of innovations, including FlashAttention for optimizing Transformer efficiency, Mamba for linear-time sequence modeling, and DPO (Direct Preference Optimization) for streamlining the alignment process. The academic community has also played a key role in developing evaluation benchmarks and enhancing model security.
- The rapid progress of open-source models has compelled the closed-source side to accelerate its own iteration cycle, while simultaneously offering developers and enterprises more autonomous and customizable model options. Chinese companies are also advancing open-source and self-reliant alternatives, including Alibaba's Qwen3, DeepSeek's V3 and R1, and MiniMax's Text-01 and M1.

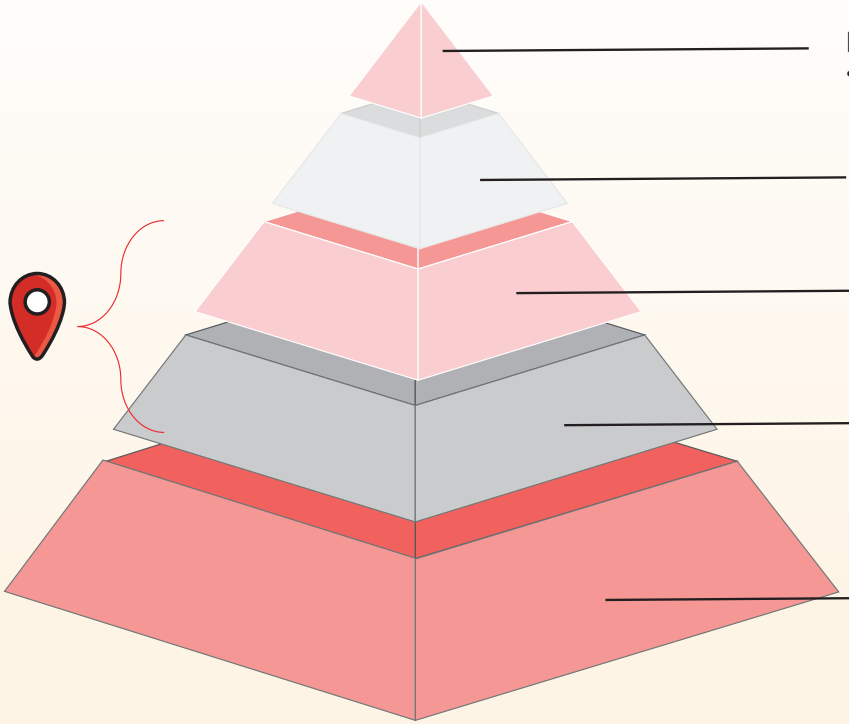
Source: China Insights Consultancy

Current foundation models are between reasoners and agents and the development path clearly indicates a rapid advancement toward AGI.

Position of current models in OpenAI's five-tier roadmap toward AGI

Key analysis

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



Level 5 Organizations

- Organizations are AI systems capable of executing the collective intelligence of an entire organization. These systems manage complex workflows and oversee multiple projects at scale across functions and departments, handling a wide range of tasks.

Level 4 Innovators

- Innovators can autonomously drive innovation in specialized fields such as science and medicine, creating novel solutions to existing problems and addressing challenges once thought to be unsolvable.

Level 3 Agents

- Agents exhibit greater autonomy, as they take initiative to perform tasks and make decisions on behalf of users, shifting from passive responders to proactive agents, without the need for human supervision.

Level 2 Reasoners

- Reasoners possess problem-solving abilities comparable to those of a PhD holder. Although lacking specialized tools, they can handle tasks that require cross-domain logical reasoning and complex contextual understanding.

Level 1 Chatbots

- Chatbots are proficient in dialogue tasks, capable of understanding natural language, engaging in conversations in a human-like manner, and executing tasks based on user instructions

Source: China Insights Consultancy

Foundation models are continuously expanding modalities.

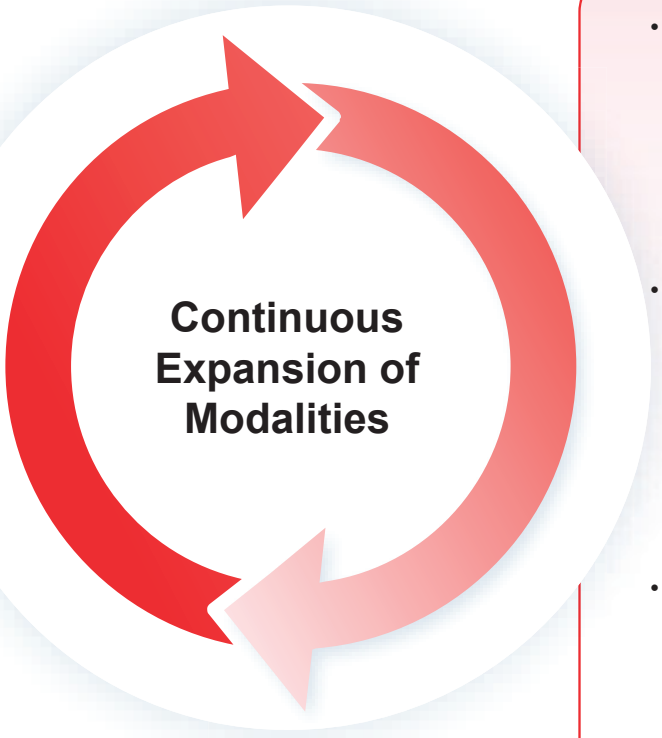
From single-modal to multi-modal

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

Visual understanding

Audio generation

Visual generation



Unified multi-modal understanding and generation

- In recent years, the academic community has begun exploring unified models capable of both multi-modal understanding and generation. These models are designed to handle diverse input modalities and generate outputs across one or more of those modalities within a single, cohesive architecture.
- Such unified systems need to combine the advantages of autoregressive models in reasoning and text generation, with the robustness of diffusion models in high-fidelity image generation. Current unified multi-modal models generally follow one of three architectural paradigms: diffusion-based, autoregression-based, or a hybrid.
- This pursuit of integration mirrors the deeper nature of human intelligence – human understanding and expression, as well as inputs and outputs across different modalities, are deeply intertwined and inseparable, rather than being modular and independent of each other.

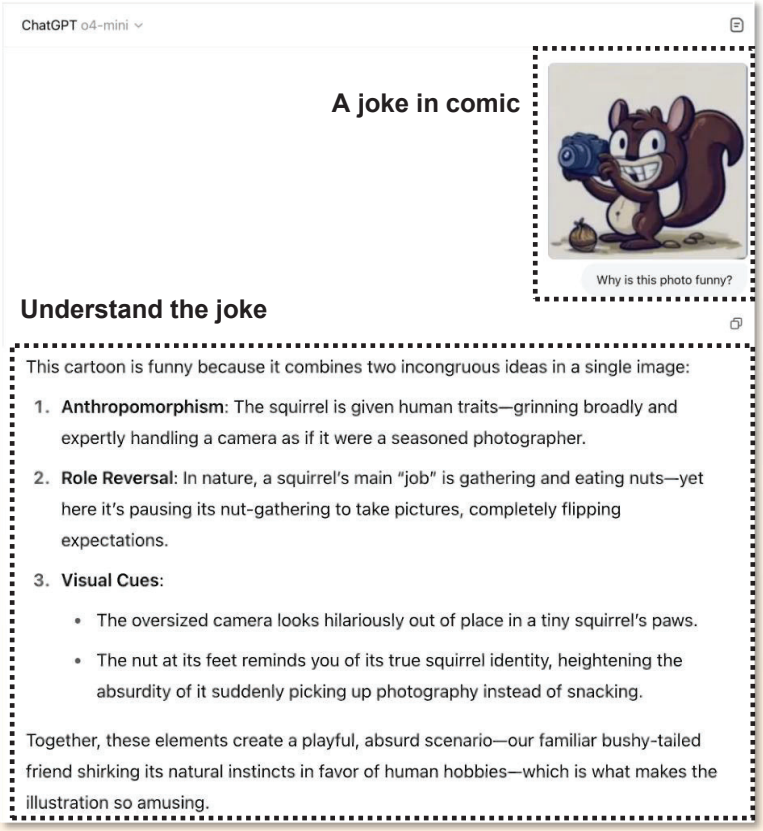
Source: China Insights Consultancy

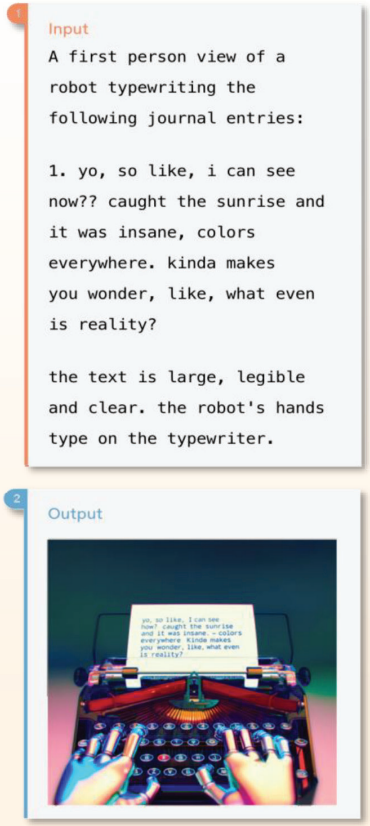
From single-modal to multi-modal (1/3): Visual understanding

Visual understanding

Key analysis

- In the early stages of multi-modal understanding, models like CLIP, ViLBERT, and VisualBERT primarily relied on dual-encoder architectures to align visual and textual inputs. More recently, the trend has been shifting towards more unified multi-modal capabilities. GPT-4V, for example, extends the GPT-4 framework to support image inputs, allowing users to ask the model to analyze visual content, describe image details, interpret humor in memes and information in medical images. Built on a decoder-only architecture, Gemini supports image, video, and audio modalities, with Gemini Ultra setting new benchmarks in multi-modal reasoning tasks.





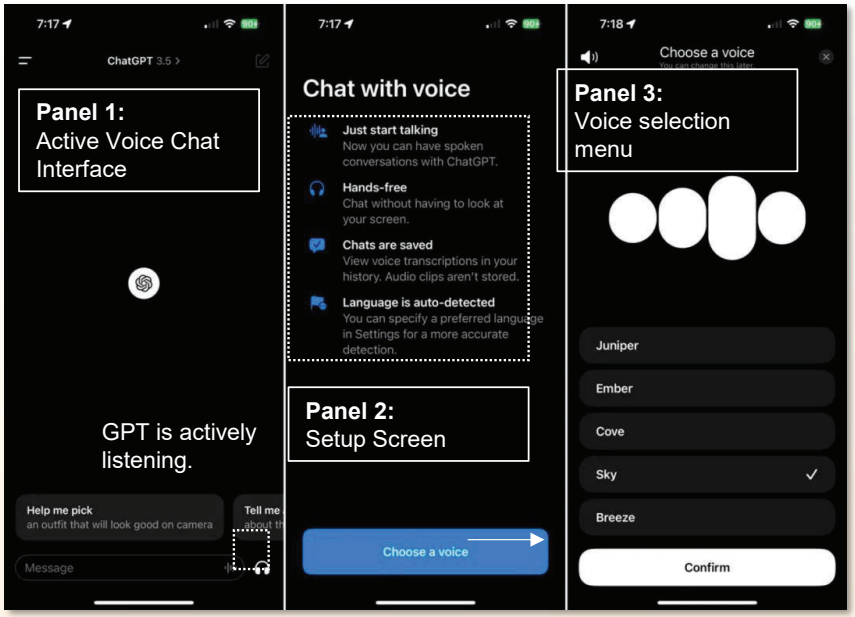
Source: China Insights Consultancy

From single-modal to multi-modal (2/3): Audio generation

Audio generation

Key analysis

- The integration of text and audio allows AI to interpret and generate audio itself.
- Audio recognition technology has made breakthroughs with the help of foundation models. OpenAI's Whisper, an open-source model released in 2022 with 1.6 billion parameters, can transcribe and translate audio in 97 languages, achieving near-human-level accuracy on English transcription. Whisper allows developers to seamlessly convert audio inputs into text for further processing by foundation models.
- Audio synthesis technology has similarly advanced at a rapid pace. In 2023, service providers such as ElevenLabs and MiniMax emerged, enabling models to read any text in natural, human-like voices. In the same year, OpenAI added audio-based conversational capabilities to ChatGPT, which is then able to interpret spoken questions in real time and respond with synthesized speech. This type of "listening-and-speaking" chatbots pave the way for broader applications of AI such as intelligent assistants and in customer service.
- The convergence of audio and text modalities has also spawned new product forms, such as voice-driven AI assistants and foundation model services embedded in smart speakers. Looking ahead, foundation models are expected to develop a deeper understanding of emotion and intent in spoken language, and generate more realistic audio responses, making human-machine interaction more natural and efficient.



Speech-To-Text → ChatGPT → Text-To-Speech

From single-modal to multi-modal (3/3): Visual generation

Visual generation

Key Analysis

- By around 2022, the output quality of leading text-to-image models began to approach that of real photographs and human-created artwork. Key examples include OpenAI's DALL-E, Google Brain's Imagen, Stability AI's Stable Diffusion, and Midjourney.
- These models typically rely on latent diffusion models, combining a language model that converts text inputs into latent representations with a generative model that produces images based on these representations. Within this process, text encoding often employs the Transformer architecture, while model training relies on large-scale image-text paired datasets.
- Within the typical models, DALL-E 3 excels in natural language interactions and is capable of generating and editing in-image text based on user prompts. Stable Diffusion is renowned for its photo realism, high degree of customizability, and active open-source community. Midjourney stands out for its artistic stylization and ease of use.
- Video, as sequential visual data, has emerged as a new frontier of multi-modal AI since 2023. With the maturing of image models, technological research has extended to video. OpenAI's Sora, a DiT-powered video model, can generate new video content from inputs in the forms of text, image, or even video. Other offerings, such as Hailuo AI and Google Veo 3, have also gained global traction. These tools have democratized creative content generation and improved workflow efficiency in the creative industry.
- The industry is now exploring auto-regressive approaches – to improve model instruction adherence and content editing capabilities by integrating DiT models with auto-regressive approaches employed in foundation models.

DALL·E

Stable Diffusion

Midjourney

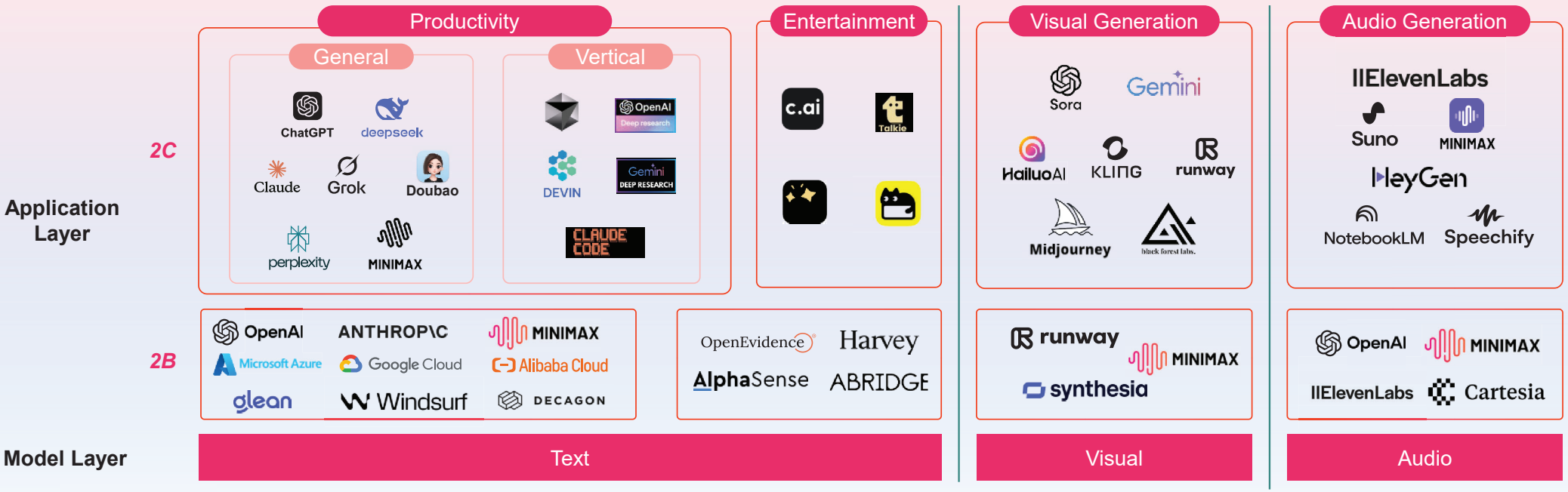
- **DALLE·3** excels at **natural language understanding**, particularly in generating images with accurate interpretations of text prompts.
- **Stable Diffusion** is renowned for its **photo-level** realism and customizability, supported by an active **open-source** community.
- **Midjourney** stands out for its **artistic style and ease of use**, popular among **creative** users.

Prompt	Output video
• Dynamic motion, 30x speed. • Camera follows a translucent white plastic grocery bag with bold red letters printed on it that read "THANK YOU" as it flies organically in the wind of a desert. • The slightly opaque bag undulates in the wind, maintaining the bold red "THANK YOU" text printed on it.	

Source: China Insights Consultancy

Currently, major applications of foundation models include productivity, entertainment, visual generation, audio generation and general 2B services.

Market landscape of foundation model applications



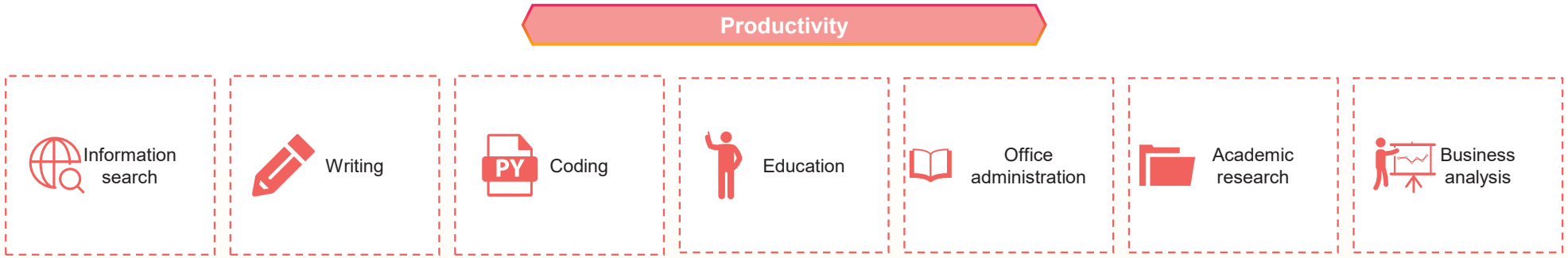
Key Analysis

Across these segments, the generalization capabilities of foundation models are enabling both large-scale deployment and personalized applications through a single, highly scalable architecture—serving a broad spectrum of user needs, from large enterprises to individual creators, and delivering strong model ROI.

Products that can achieve sustained organic growth over time are those driven by continuous improvements in underlying model intelligence. Players that intend to maintain long-term market leadership must develop proprietary foundation models that can be optimized from end-to-end and consistently maintain top-tier performance. A breakthrough in foundation model intelligence can rapidly propel them to prominence, as users naturally gravitate towards technologies that offer a markedly superior experience.

Source: China Insights Consultancy

The productivity segment represents a massive opportunity with broad downstream use cases, where users are often engaged with multiple products simultaneously.



Key Analysis

The productivity segment represents a massive opportunity with broad downstream use cases. Users often engage with multiple models simultaneously. Top use cases include information search, writing, coding, education, office administration, academic research, and business analysis, which cover all aspects of work and daily life.

The productivity segment has undergone a generational shift from chatbots in 2023 to agents in 2025. Leading chatbot players include ChatGPT and DeepSeek R1, with OpenAI Deep Research and MiniMax Agent driving the next generation of AI agents. Unlike chatbots that simply respond to prompts, agents can complete long-horizon tasks due to advances in multi-step reasoning and use of external tools, enabling them to learn and improve through interactions with their environment. Long-term leaders in this field must possess end-to-end capabilities, the ability of models to enhance their capabilities via end-to-end reinforcement learning using proprietary models and rewards from application-specific environments.

Since the second half of 2024, AI coding applications experienced exponential growth, fueled by the breakthrough of Claude 3.5 Sonnet. Its capabilities in code design, debugging, and optimization have powered over 30 million developers worldwide. Notable products include Cursor, a code editor for professional developers, and Windsurf, an enterprise-level secure coding platform. The overall trend is shifting from simple code completion towards more advanced coding agent capabilities, with a long-term potential for enabling full-stack personalized software generation from a single chat interface. This would not only lower barriers for professional product development, but also unlock a new market for users with little experience.

Source: China Insights Consultancy

Entertainment is the second-largest segment following productivity.

Entertainment

10M
young users worldwide

Long-tail

create & interact with **personalized AI agents**

- Spanning role-play
- Companionship
- A wide range of everyday Q&A interactions

Competition in the entertainment segment is stabilizing.

character.ai

MINIMAX

Key Analysis

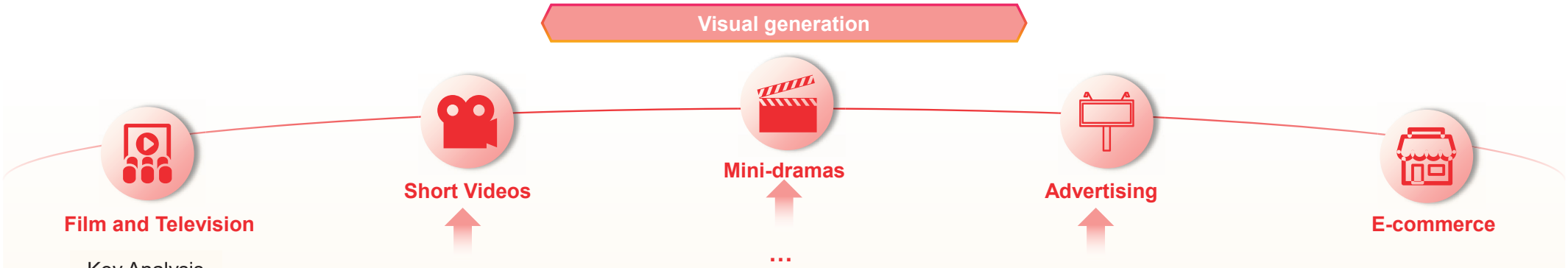
Entertainment is the second-largest segment following productivity, with tens of millions of young users worldwide creating and interacting with personalized AI agents. The use cases are highly diverse, spanning role-play, companionship, and a wide range of everyday Q&A interactions.

Competition in the entertainment segment is in the process of stabilizing. Leading products include Character AI and Talkie / Xingye, which can enhance user experience with model optimization and inspire users' creativity with rich multi-modal creative tools. This combination drives high engagement, strong user stickiness, and highly interactive experiences.

The new generation of AI-native users are naturally inclined to interact with AI companions. As societal productivity continues to rise and material needs are increasingly met, entertainment AI products will tap into users' emotional and psychological needs and unlock long-term market potential through personalized, emotionally resonant experiences.

Source: China Insights Consultancy

Image generation has emerged as the first AI domain to achieve commercialization.



Key Analysis

In 2022, models like Midjourney impressed the world with their visually striking outputs, sparking exponential growth in social media engagement as image quality progressively met commercial standards. The primary user base consists of professional creators and enthusiasts in graphic design, film/TV production, advertising and e-commerce. These tools not only inspire creativity but also significantly enhance design workflow efficiency.

Leading applications in this space, including GPT-4o, Midjourney and Flux, continue to make breakthroughs in image quality, editing customizability, and diversity of styles. These advancements have unlocked greater end-user application scenarios, such as professional-standard product design and commercial marketing materials. AI-generated images have already achieved widespread popularity, marking their evolution from being just creative tools to becoming mainstream content.

Video generation has emerged as a rapidly growing segment in 2024, with a clear product-market fit. Demand comes from a wide range of industries, including film and television, short videos, mini-dramas, advertising, and e-commerce, leading to a massive market opportunity. In these industries, conventional video production often requires an entire team, whereas foundation models open up new market opportunities for individual professional creators to act as “one-person studios” to produce high-value content as well as enhancing their productivity.

Leading players in this segment include Sora, Hailuo AI, and Kling, among others. Their core competitiveness lies in maintaining cutting-edge model R&D capabilities and cost efficiencies, coupled with fast-iterating creative workflow features and a vibrant creator ecosystem.

AI-generated videos are beginning to go viral increasingly frequently on social media, signaling that model performance is beginning to break through the boundaries of consumer-level content. In the future, relevant products may evolve into a “real-time personalized video generation engine”, lowering barriers for anyone to create and consume personalized content.

Source: China Insights Consultancy

Audio generation and general 2B services are rapidly expanding application areas for foundation models.

Audio generation

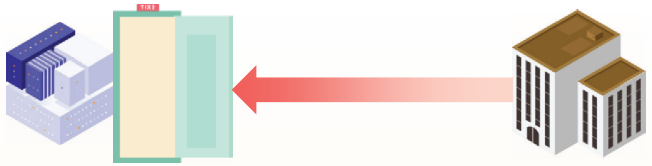
Key Analysis

Audio is the universal interface of interaction in the AI era, with a broad downstream application market. For enterprises, AI voice agents overcome the limit of human capacity in sales and customer service, including recruitment, finance, healthcare; for content creators, it enables lifelike and emotionally expressive audio generation for audiobooks, education, dubbing and gaming, and others.

Leading players include OpenAI, MiniMax, and ElevenLabs. Their core competitiveness lies in delivering hyper-realistic audio model quality while maintaining low cost and low latency

Numerous agentic AI applications, and smart devices are empowered by audio in the AI era. OpenAI's GPT-4o introduced real-time audio interaction in May 2024, setting a new standard for chatbots; Google's NotebookLM saw viral success in September 2024 with its podcast generation feature. As human–AI interactions grow exponentially, the audio submarket holds vast untapped potential.

General 2B services



Key Analysis

To accelerate AI adoption in various fields, foundation model companies such as OpenAI and Anthropic typically offer model capabilities to developers and enterprise clients via APIs with an open-platform strategy. Cloud service providers such as Microsoft, Amazon, Google, and Alibaba also provide models, toolkits and professional services through APIs, industry-tailored solutions, and on-premise deployment.

The core competitiveness in this segment includes model performance, cost-efficiency, and stability during high concurrencies, which are the top concerns for developers and enterprise customers. Secondary considerations include security, compliance, and customer support. A multiple-model strategy is now common, with enterprises often using three or more models and routing different models to specific tasks based on use-case requirements.

Enterprise demand is surging across industries. As agentic models become increasingly capable of delivering satisfying outcomes and inference costs continue to drop rapidly, foundation models are set to become a new productivity norm, continuously unlocking value across sectors.

Source: China Insights Consultancy

Table of Content

The global foundation model market comprised of revenue generated by model-based and deployment-based approaches.

Overview of foundation model

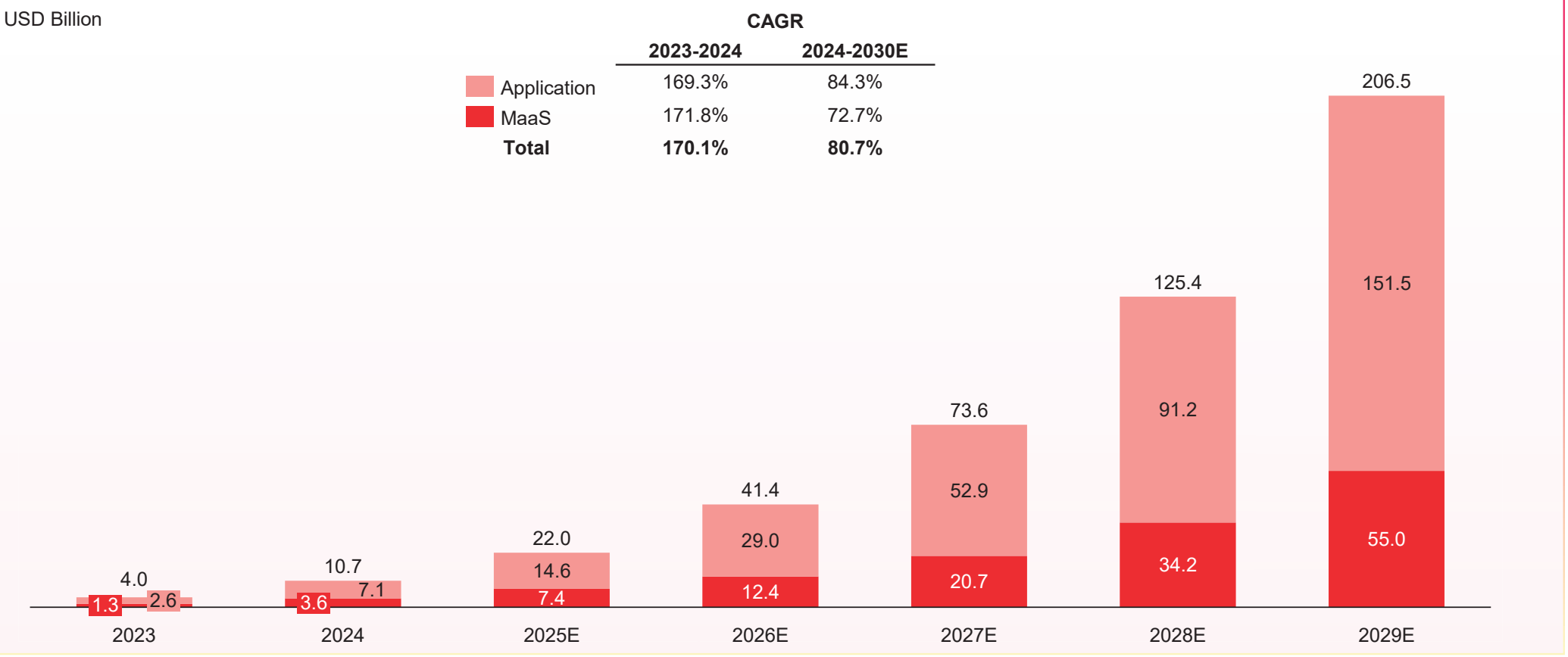
- Foundation model technology remains in a stage of rapid development. Compared to a deployment-based approach, the model-based approach allows users to benefit from continuous model improvements without incurring version migration costs. Users can also dynamically scale their model usage based on actual demand, reducing upfront investments and ongoing maintenance expenses related to hardware and infrastructure. Moreover, this approach supports automatic resource scaling to meet users' evolving needs.

Customer & Revenue streams		Definition	Pricing model	Specific use cases	Business model
APP	Consumer subscriptions	• Users pay ***recurring fees for tiered access*** to services	• Monthly subscriptions***(e.g., $20/month)***; • Feature-based tiers (Basic/Pro)	• Users pay a monthly fee to access premium features such as ***faster response, advanced reasoning, or priority support***.	• ***Cloud-based sevice***
APP	Online marketing services	• Monetization via user traffic (free service + ***ads/data insights***)	• Advertiser pays (CPM/CPC); • Anonymized data licensing	• Users access free model services while being shown ***targeted ads*** based on their interaction data.	• ***Cloud-based service***
APP	Enterprise subscriptions	• Pay-as-you-go access to model capabilities for developers	• Per token/request ***(e.g., $0.002/1k tokens);*** • Volume discounts	• Developers integrate model capabilities into their own products, ***paying per token or per query.***	• ***Cloud-based service***
MaaS	Cloud-based API calls	• ***Pre-built*** industry SaaS products or plugins (ready-to-use)	• Per-user/per-module subscription. ***(e.g., $50/user/month);*** • One-time license	• Businesses subscribe to ***ready-made tools powered by LMs*** for writing, summarizing, or customer interaction automation.	• ***Cloud-based service***
Model-based revenue	On-premise deployment	• Full model deployment on client's servers with complete control	• Upfront deployment fee ***($500k+)***; • Annual maintenance ***(15-30% of contract)***	• Enterprises deploy foundation models on ***internal servers*** to ensure data security and full system control.	• On-premise deployment service

Source: China Insights Consultancy

The global model-based foundation model market is still in the early stages of commercialization.

Market size of foundation model industry, in terms of model-based revenue, global, 2020-2030E



Source: China Insights Consultancy

Drivers of global foundation model industry

Drivers of global foundation model industry

1 Technological leaps

- The foundation model market is characterized by disruptive technological breakthroughs, with the improvements in each new generation of foundation models expanding the scope of potential applications.

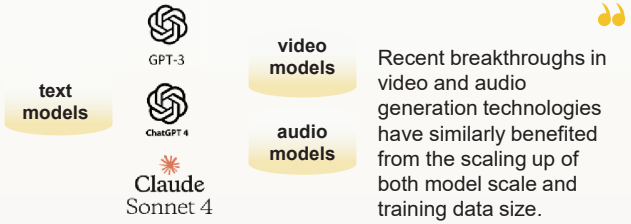
GPT-3	enabled casual **chatbots** to first achieve product-market fit
GPT-3.5	brought chatbot applications to a highly usable level
GPT-4	tapped into **professional domains** such as finance and law
Sora	drove video generation to meet the commercial requirements for quality
GPT-4o	facilitated a new surge (**multi-model**) in new user adoption for ChatGPT

- Technology serves as the most powerful underlying driver to the wave of foundation models, with foundation model companies positioned at the beginning of this transformative growth. In the early stages of each new generation of models, creative use cases emerge naturally in response to the enhanced model capabilities, which are initially long-tailed and scattered but would gradually converge into several mainstream product-market fits. While models within the same generation may experience performance plateaus, true breakthroughs often occur between generations, with each leap climbing a new technology curve and unlocking new application scenarios.

2 Scaling Law

- The fundamental driver behind market growth of foundation models lies in the fact that the foundation model technology is able to keep scaling up.

The pre-training scaling law is now democratized.
The model performance improves in proportion to the increases in model scale, data size and computing power.



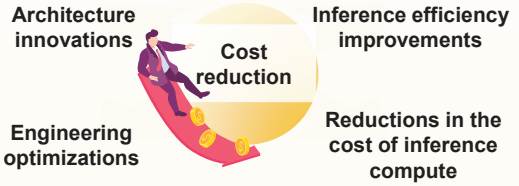
Recent breakthroughs in video and audio generation technologies have similarly benefited from the scaling up of both model scale and training data size.

- Beginning with OpenAI's o1, the industry witnessed a new scaling law focused on test-time compute. The model's reasoning capability enhances as the computational load in inference extends – the longer the time that the model spends on thinking, the better the performance. This principle has been consistently validated in subsequent models such as OpenAI's o3.
- Looking ahead, the scaling up of foundation models is expected to continue, with the scaling of both pre-training and inference reinforcing each other. This dynamic is driving exponential growth in model performance and usage, and is expected to sustain rapid market expansion.

3 Cost reduction

Declining model cost
Improvements in model capabilities
→ **Unlock an increasing number of use cases**
✓ **Cross the ROI threshold**
✓ **Achieve product-market fit**

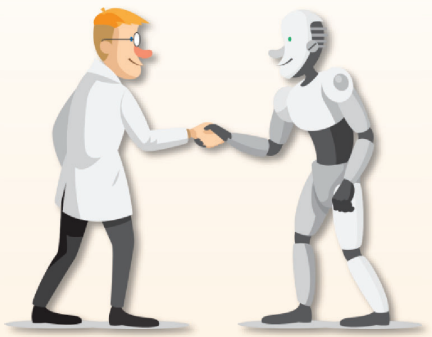
- At the time of GPT-4's release, many vertical applications, such as content moderation, already met performance requirements but remained commercially unviable due to high costs and negative ROI. The inference cost of foundation models has been decreasing steadily, with the per-token cost dropping by over 99% since the release of GPT-4, enabling broader adoption across high-volume, backend industry scenarios. This decline has been driven by a combination of architecture innovations, inference efficiency improvements, engineering optimizations, and reductions in the cost of compute. These factors are expected to continually lower costs at a predictable rate.



Source: China Insights Consultancy

Trends of global foundational model applications (1/3)

1 Agent applications



Agents are capable of operating at a professional level, acting autonomously, and delivering end-to-end results, ultimately driving GDP of trillions of dollars' worth

- Achieving professional-level performance involves the execution of specialized tasks within expert domains. Acting autonomously allows systems to operate independently of human time and attention. Delivering end-to-end results signifies the ability to generate economic value. These characteristics mark the inflection point for LLMs transitioning from offering tools to delivering results. Consequently, the addressable market for LLMs will expand beyond the boundaries of enterprise software budget and into the broader market space for labor services.
- In addition, agents are continuously enhancing their capabilities to complete tasks. Agents also follow its own scaling law –with the duration of tasks that agents can autonomously handle doubling approximately every seven months. Today, AI can autonomously complete tasks that typically take humans one hour to complete. AI is expected to autonomously handle 2-3-hour tasks by 2026, and one-month tasks within five years. This paves the way for a future of "infinite experts" – AI software engineers, financial analysts, and research scientists contributing to greater productivity and agent economy, 24/7 without downtime.
- In light of the agent scaling law, future agents will not only execute tasks, but also act as AI researchers that are able to accelerate the research of new proprietary algorithms and develop new agents that surpass their own capabilities. This exponential, self-reinforcing evolution is precisely what sets this generation of AI technology apart at its core. A compounding effect will emerge between the scaling of algorithm and application, each reinforcing and accelerating the other. This feedback loop could make AI the fastest-moving technological revolution in human history.

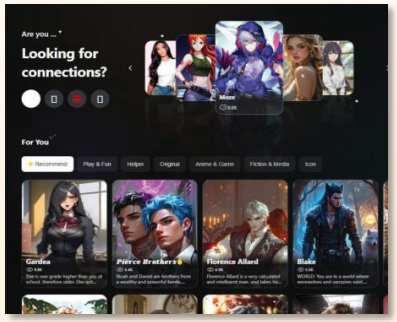
Source: China Insights Consultancy

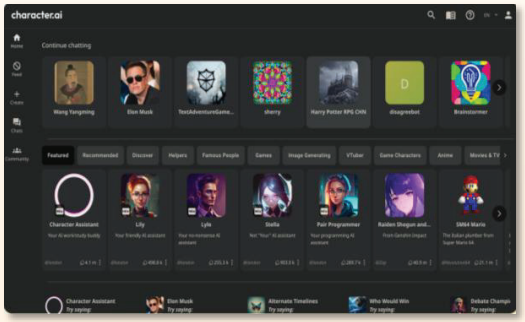
Trends of global foundational model applications (2/3)

2 Entertainment and generative applications

Entertainment and generative applications: rapid growth across multiple verticals

- The new generation of AI native users seek immersive co-creation experiences with AI, driving entertainment products to evolve toward personalized emotional companionship. The evolving role of AI is creating opportunities for hybrid virtual-physical social and entertainment scenarios. Advances in video generation are altering the limits of creativity, as AI-generated videos have the potential to become viral on social media. This indicates a shift in content production from professional tools to widely accessible creative engines, significantly changing the video content supply landscape. Meanwhile, audio generation and interaction capabilities are gradually becoming standard across AI applications, advancing from basic command-response functions toward more emotionally expressive communication. This evolution is enabling more natural interactions between human and machines, positioning the voice interface as a central hub for multi-modal interaction.



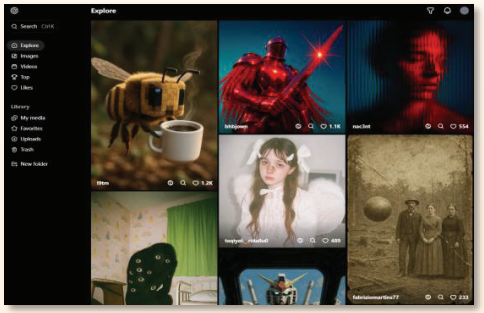


Source: China Insights Consultancy

Trends of global foundational model applications (3/3)

3 Multi-modal applications

Multi-modal applications: integration of modalities unlocks new market potential



- In recent years, the market for visual understanding applications has grown rapidly. For example, as the capabilities of visual understanding models improve, token consumption in K-12 online education has surged. Use cases in traditional industries, such as intelligent inspection and video search, are also rapidly adopting next-generation visual models to power new applications.
- In March 2025, the release of GPT-4o introduced image generation capabilities, significantly improving image quality and triggering a sharp uptick in new ChatGPT subscriptions, demonstrating the strong commercial potential of multi-modal integration. Unlike previous approaches that relied on separate models like DALL-E for image generation, GPT-4o is built on a natively multi-modal architecture that generates image directly from text prompts. This has brought a new level of precision and control to image generation.
- The ability to accurately render text within images and to edit images with fine-grained control has opened up commercial use cases such as generating educational visuals, product posters with stylized typography, and scientific illustrations. Looking ahead, deeper integration of video, audio, and text modalities will make it possible to create fully editable videos, generate synchronized audio and text along with the video content, and more, unlocking market opportunities on the scale of the short-form video revolution.

Source: China Insights Consultancy

Table of Content

MiniMax is the tenth largest foundation model technology company globally, and the second largest non-U.S. player, in terms of model-based revenues in 2024.

Ranking of global foundation model technology companies, in terms of model-based revenues in 2024

Rank	Company name	Market share, %	Rank	Company name	Market share, %
1	Company A	30.1%			
2	Company B	16.9%	9	Company I	0.3%
3	Company C	8.2%	**10**	**MiniMax**	**0.3%**
4	Company D	4.7%	11	Company J	0.3%
5	Company E	2.8%	12	Company K	0.3%
6	Company F	1.8%	13	Company L	0.3%
7	Company G	0.7%	14	Company M	0.2%
8	Company H	0.5%	15	Company N	0.2%

Source: China Insights Consultancy

MiniMax is the fourth largest pureplay foundation model technology company globally, and the largest non-U.S. player, in terms of model-based revenues in 2024.

Ranking of global pureplay foundation model technology companies, in terms of model-based revenues in 2024

Rank	Company name	Market share, %
1	Company A	30.1%
2	Company D	4.7%
3	Company E	2.8%
4	**MiniMax**	**0.3%**
5	Company J	0.3%

Source: China Insights Consultancy

MiniMax's foundation models have achieved leading performance across text, video and speech modalities (1/3).

Artificial Analysis Intelligence Index (evaluation of text models), November 7, 2025

Rank	Company name	Model	Index
1	OpenAI	GPT-5 Codex (high)	68
1	OpenAI	GPT-5 (high)	68
3	X	Grok 4	65
4	Anthropic	Claude 4.5 Sonnet	63
5	**MiniMax**	**MiniMax M2**	**61**
5	OpenAI	gpt-oss-120B (high)	61
7	X	Grok 4 Fast	60
7	Google	Gemini 2.5 Pro	60
9	Anthropic	Claude 4.1 Opus	59
10	Alibaba	Qwen3 235B A22B 2507	57

Source: Artificial Analysis, China Insights Consultancy

MiniMax's foundation models have achieved leading performance across text, video and speech modalities (2/3).

Artificial Analysis Speech Arena Leaderboard (evaluation of speech models), June 22, 2025

Rank	Company name	Model	Arena ELO
1	ByteDance	Seedance 1.0	1,355
2	**MiniMax**	**Hailuo 02**	**1,331**
3	Google	Veo 3 Preview (No Audio)	1,244
4	Kuaishou	Kling 2.0	1,195
5	Kuaishou	Kling 1.6 (Pro)	1,144
6	Runway	Runway Gen 4	1,120
7	Google	Veo 2	1,118
8	Lightricks	LTV Video v0.9.7 (13B)	1,064
9	MiniMax	I2V-01-Director	1,047
10	Runway	Runway Gen 3 Alpha Turbo	1,005

Source: Artificial Analysis, China Insights Consultancy

MiniMax's foundation models have achieved leading performance across text, video and speech modalities (3/3).

Artificial Analysis Video Arena Leaderboard (evaluation of video models), June 22, 2025

Rank	Company name	Model	Arena ELO
1	**MiniMax**	**Speech-02-HD**	**1,174**
2	OpenAI	TTS-1 HD	1,146
3	OpenAI	TTS-1	1,132
4	ElevenLabs	Multilingual v2	1,114
5	ElevenLabs	Turbo v2.5	1,108
6	Cartesia	Sonic English (Oct'24)	1,103
7	Kokoro	Kokoro 82M v1.0	1,078
8	Microsoft	Azure Neutral	1,056
9	Amazon	Polly Long-Form	1,056
10	Google	Studio	1,039

Source: Artificial Analysis, China Insights Consultancy

Competitive barriers

1 R&D capabilities of foundation models

- The competitiveness of foundation model products is fundamentally based on the underlying foundation models. Performance improvements driven by the iteration of foundation models often far outweigh enhancements made at the application layer or through product refinement. As a result, leading foundation model products nowadays are typically developed by companies with strong in-house foundation model R&D capabilities, while users tend to gravitate toward top-tier products that offer the best experience. Given the rapid pace of technological advancement, players in the industry must continue investing heavily in R&D to maintain performance leadership and their competitive edge.

2 Commercialization capabilities

- Strong commercialization capabilities enable foundation model companies to translate cutting-edge research and technologies into usable products more rapidly, shortening the cycle from technological development to tangible commercial value. By strategically selecting and developing products with the greatest potential for scalable commercialization, foundation model companies can further amplify the market impact of technological breakthroughs, improve the ROI of model development, and support the long-term sustainability of ongoing research efforts.

3 Exceptional organizational abilities

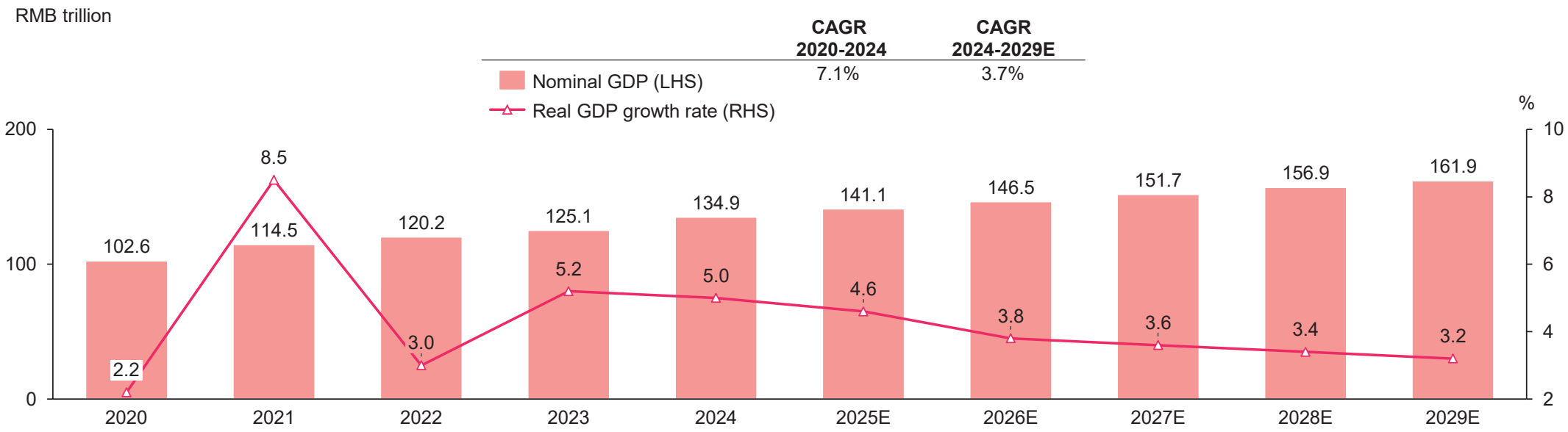
- Developing foundation models spans multiple complex domains, including advanced algorithms, AI infrastructure efficiency improvement, and data governance, among others. Building an organization capable of attracting top-tier talents is critical to overcoming technical bottlenecks. A culture that fosters innovation and passion attracts and retains world-class AI talents, laying the foundation for a long-term, sustainable competitive advantage.

Source: China Insights Consulting

Table of Content

China's real GDP growth rate witnessed a sharp decline in 2020 and 2022, however, it is expected to stabilize in the future, fueled by increasing consumer demand and an opening-up policy.

Nominal GDP and Real GDP growth rates, China, 2020-2029E



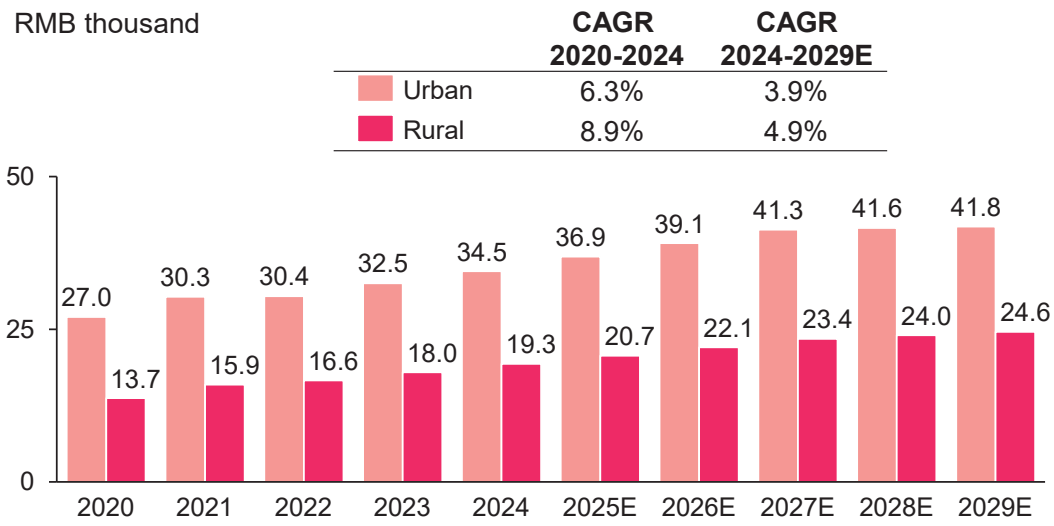
Key analysis

- China’s economy has shown significant growth in recent years, with its nominal GDP rising from RMB 99.1 trillion in 2019 to RMB 134.9 trillion in 2024 and is projected to reach RMB 161.9 trillion by 2029. However, the value of real GDP growth took a plunge in 2020 due to the covid-19. Although there was a one-time increase in real GDP during the economic recovery in 2021, the future GDP growth is expected to remain stable after 2023.
- China's economic outlook is currently facing several challenges. The ongoing trade tensions with the United States have led to uncertainty and a potential slowdown in economic growth. In addition, China's aging population is putting pressure on its labor force and social security system. Despite these challenges, China continues to invest heavily in infrastructure and technology, which could drive future economic growth. The government's commitment to economic reform and opening up to foreign investment also presents opportunities for sustained growth.

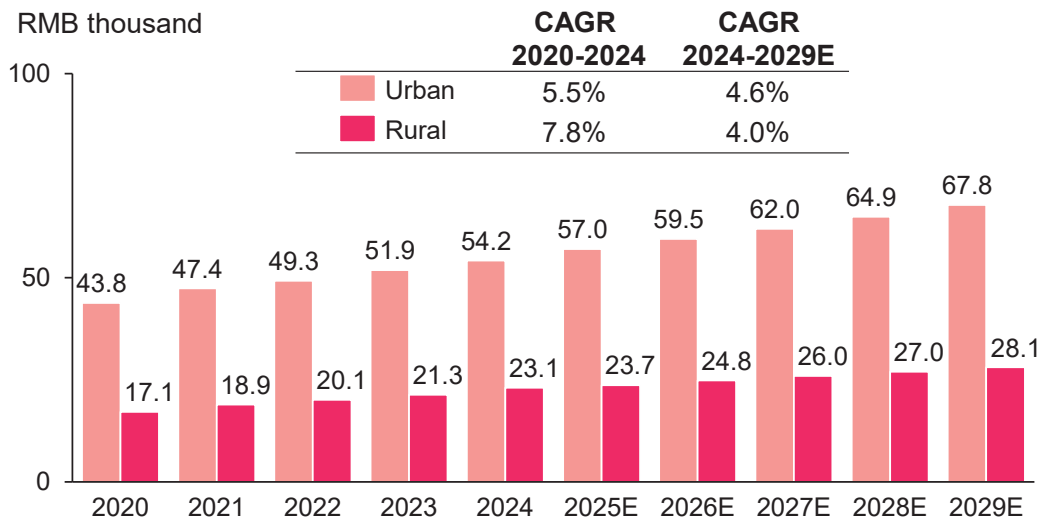
Source: NBS, China Insights Consultancy

Urban and rural per capita consumption rose; despite urbanization shifts, higher incomes drive increased consumer spending.

Per capita consumption expenditures in China, urban and rural, 2020-2029E



Per capita disposable income in China, urban and rural, 2020-2029E



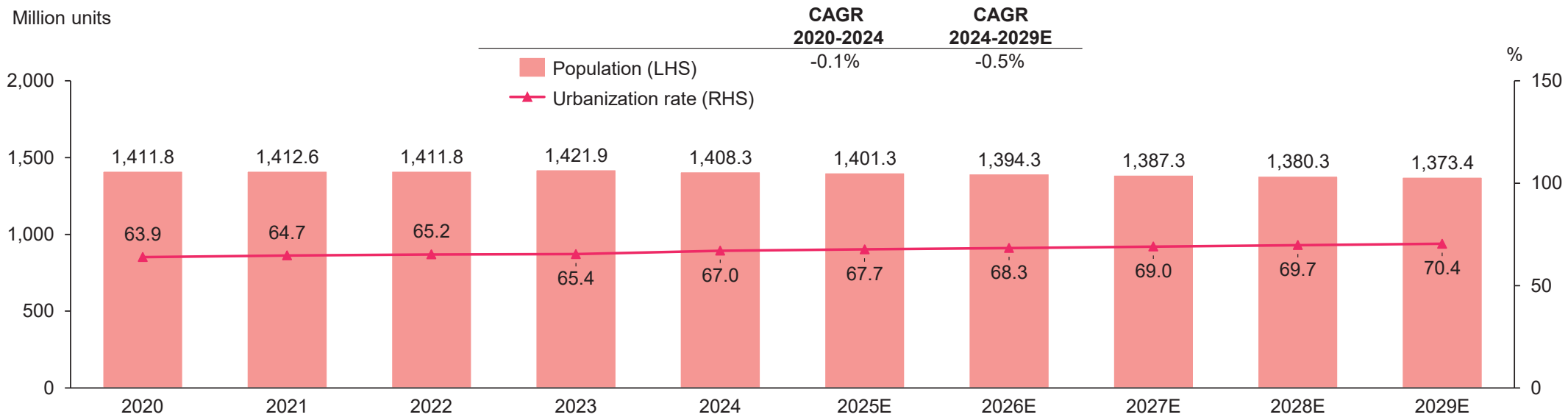
Key analysis

- In 2024, the country‘s per capita consumption for urban and rural areas showed growth, with urban resident’s spending increasing nominally by 4.7%, reaching around RMB 34.5 thousand, while rural resident’s consumption grew by 6.1%, surpassing RMB 19 thousand over the same period. The national per capita disposable income of urban residents increased by 4.6% to RMB 54.2 thousand in 2024, while rural residents saw a 6.6% increase to reach RMB 23.1 thousand.
- The ongoing increase in consumption expenditure is the direct result of rising per capita disposable income. This growth in income has mainly been driven by the rapid development of the Chinese economy and the continuous upgrading of China's manufacturing sector, along with a structural shift in the mainstay of development, from the primary and secondary sectors to the tertiary sector. Growing purchasing power and consumption upgrades in both urban and rural areas suggest a stronger level of consumer confidence and sustainable internal circulation, thus giving rise to the new retail and logistics industry. It provides tremendous development potential for China’s consumer and FMCG industries.
- The compound annual growth rates of per capita disposable income and consumption expenditure in the urban area is expected to decline in the future due to lower birth rates and an ageing population. Despite this, disposable income and expenditure levels will still grow steadily, driven by the globalization of trade, productivity gains and the internationalization of the Renminbi.

Source: NBS, China Insights Consultancy

Despite the low population growth due to the one-child policy, China's population size remains globally significant.

Population and Urbanization, China, 2020-2029E



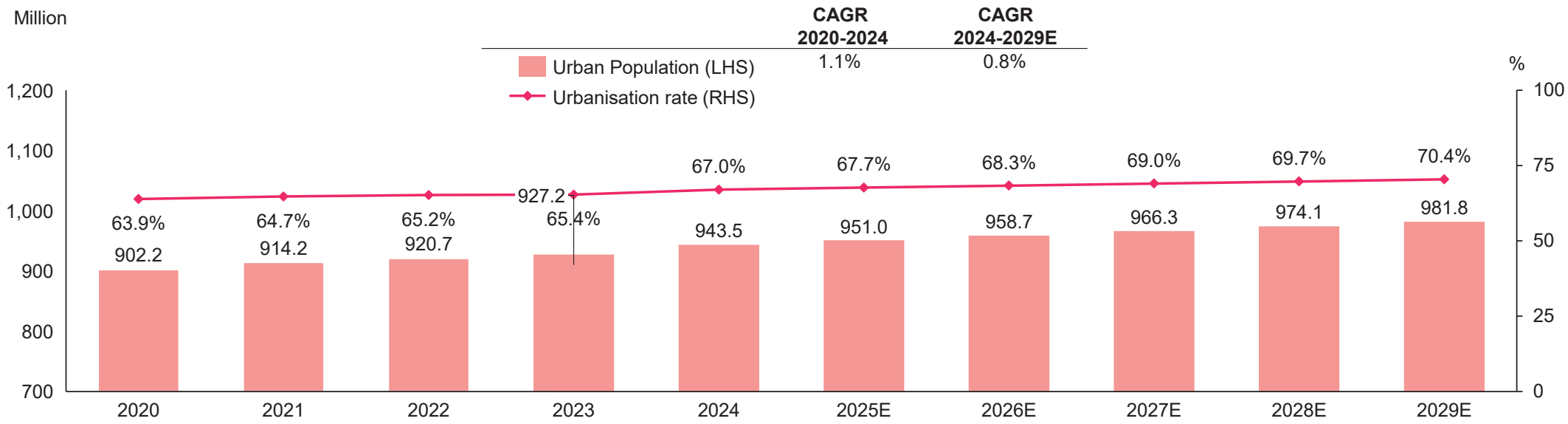
Key analysis

- China's population and urbanization data highlight the intricate nature and diverse aspects of its economic and social progress. Being the second most populous country, accounting for 20% of the global population, China is confronted with both challenges and prospects due to its enormous size. Despite having a sizable population, the impact of the one-child policy has led to consistently low population growth in recent years. It is projected China's population will reach 1.37 billion with a minus compound annual growth rate from 2020 to 2024 continuing to the 2029. By 2029, the growth rate of the urbanization rate will remain at around 70.4%.
- This suggests that more people will migrate from rural to urban areas in the coming years, fueling urban economic development but also posing challenges in areas such as urban planning, infrastructure development and social security. China's population and urbanization figures have far-reaching implications for economic and social development. As population growth continues and urbanization accelerates, the demand for resources will increase, as will the pressure on the environment.

Source: NBS, China Insights Consultancy

The urbanization rate has steadily increased, creating a substantial market for industries including digital retail.

Urban population and urbanization rate, China, 2020-2029E



Key analysis

- China presently holds the title for the world's second largest population, and this figure is expected to maintain an upward trajectory. China's population currently constitutes roughly 20% of the global population. With the advent of reform and opening up policies, urban built-up areas have experienced rapid expansion, aligning with the nation's swift urbanization. As of the end of 2024, urban residents reached 943.5 million, marking an increase from 902.2 million in 2020. At the same time, urbanization rates are expanding rapidly, regarding the narrowing gap relative to developed nations. China has already stepped into the post-industrialization period, with its urbanization rate of permanent population standing at 67.0% in 2024, up 3.1 from 2020.
- This increasing urbanization is driven by various factors, including economic opportunities in urban areas, improved infrastructure, and government policies aimed at promoting urban development. As more people migrate to cities, it presents both opportunities and challenges for China, including the need for sustainable urban planning, infrastructure development, and social welfare systems to support the growing urban population.

Source: NBS, China Insights Consultancy