HUAWEI

Striding Towards the Intelligent World White Paper 2023

Cloud Computing

Reshaping Industries with AI

Building a fully connected, intelligent world

Contents

Trend 1: AI Is Accelerating and Scaling up Across Industries

Commercial AI adoption is accelerating across industries. There are three phases to this process:

- **Point solution**: A point solution improves an existing procedure and can be adopted independently, without changing the system in which it is embedded.
- **Application solution**: An application solution enables a new procedure that can be adopted independently, without changing the system in which it is embedded.
- **System solution**: A system solution improves existing procedures or enables new procedures by changing dependent procedures.

Looking back, we see a similar pattern during the transition from steam engines to electricity. For typical steam-powered factories, the transformation first occurred on select machines, then on specific production lines, and finally on all equipment, all production lines, and entire factories.

Today, industries are **moving from point solutions to application and system solutions**, as illustrated below:

	Point solution	Application solution	System solution
Innovative practices in industries	• **Banking**: Intelligent marketing and risk control • **Telcos**: Site deployment and alarm compression • **Mining**: Foreign object identification along main conveyor belts • **Power grids**: Powerline inspection and distributed energy management • **Manufacturing**: Quality inspection • **Ports**: Container identification and security surveillance • **Airports**: One ID for airport travel • **Cement production**: Real-time optimization • **Interior design**: AI-assisted drafting • **Data centers**: Energy conservation • …	• **Banking**: Intelligent insight & risk report generation • **Mining**: Remote control of underground roadheader • **Telcos**: Field O&M assistant • **Ports**: Intelligent planning platform • **Media**: Intelligent content generation • ...	• Revamped business procedures based on technologies like **digital twins and industry-specific foundation models** (Active exploration in telecoms, finance, mining, etc.)

Source: Power and Prediction: The Disruptive Economics of Artificial Intelligence

Recommended Course of Action: Align AI Strategy with the Overall Business Strategy, Build a Solid Data Backbone, and Adapt Foundation Models to Industry Needs

AI is set to transform every corner of society. It will unleash productivity and become a core engine of growth, reshaping every industry. To accelerate this process, we will need a clear and well-executed strategy.

Business strategy + goals as the driving force

- Intelligence is a new phase of digitalization. AI strategy must be aligned with enterprises or organizations' overall strategy to support preset business goals. Any successful digital and intelligent transformation is driven by business needs rather than technologies.
- Enterprises should focus on high-value task scenarios first. AI adoption in the name of adopting new technology should be discouraged. Instead, let AI adoption be guided by the strategy and goals of the business.

Enterprise-class datasets as the foundation

- AI needs large, high-quality datasets to effectively learn and generalize patterns, make accurate predictions, and improve its performance over time. The key to scaling up AI is building an enterprise-class data backbone to aggregate data from all sources and provision clean and transparent data. Good data governance enables fast data flow within your organization and maximizes the value of your data.

Foundation models + Industries as the core

- "AI for Industries" is about bridging the chasm between AI and industry needs by adapting AI to downstream tasks. For example, BloombergGPT, Bloomberg's 50-billion parameter LLM, provides intelligent services for financial analysts. This marked the beginning of the integration between general foundation models with the finance industry.
- Foundation models are set to become the operating system of AI, as they will accelerate hardware adaptation in a cost-effective manner. All AI algorithms can be built and deployed around foundation models. Although foundation models have huge advantages in generalization and large-scale replication, they will need to dive deep into complex industry scenarios and embed extensive industry knowledge in order to create real value.

Industry Practice: Building Smart Mines with AI and an Industrial Internet Architecture

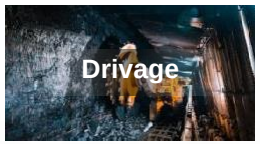
Mining is too often a dangerous undertaking. Many miners today are still working in mines that are hundreds of meters underground. The main goal of smart mines, therefore, is to improve miner safety and efficiency and reduce the underground staff (or even to run a mine with no underground staff). Smart coal mining is based on the industrial Internet architecture, where unified standards, architectures, and data specifications are applied. Under this architecture, data becomes the new oil and AI the new productivity. AI embeds industry expertise and know-how and frees miners from repetitive and dangerous tasks.

- **Point solution**: The intelligent main transport monitoring system powered by Huawei's Pangu Mining Model replaces human visual inspection. Running 24/7, this system accurately identifies anomalies like large cores and anchor rods on conveyor belts. This alone allows the coal mine to downsize the underground staff by 20%. The intelligent drivage operation monitoring system accurately monitors the compliance of drivage operations, including the drilling depth and agent stirring time. This helps to mitigate the uncertainty of human factors, improving miner safety.
- **Application solution**: 5G+AI enable new applications and working processes. Using 5G, hundreds of channels of HD videos are sent back in real time. AI stitches the videos together to create a 40-meter long, panoramic view to support high-precision remote control. This allows more coal miners to work above the ground and in a much safer working environment.
- **System solution**: Based on the industrial Internet architecture, the IoT platform collects data from 3,000+ pieces of mining equipment and stores the data in unified formats in a single data lake. A big data foundation streamlines different coal mine subsystems and unifies OT and IT data governance, providing comprehensive, real-time, all-scenario data for coal mine operators. The information can be used to quickly build a digital twin of coal mines. With the Pangu Mining Model, we can quickly replicate task-focused AI solutions across 21 task scenarios of all 9 working processes in a coal mine, including mining, drivage, main transport, auxiliary transport, lifting, security monitoring, anti-burst and pressure relief, coal separation, and coking.

AI empowers scenario-specific intelligence

1.0 Point solution

Main transport: 98% accuracy in foreign object recognition
Drivage: 95% accuracy in operational compliance detection.



5G+AI enable new applications and processes

2.0 Application solution

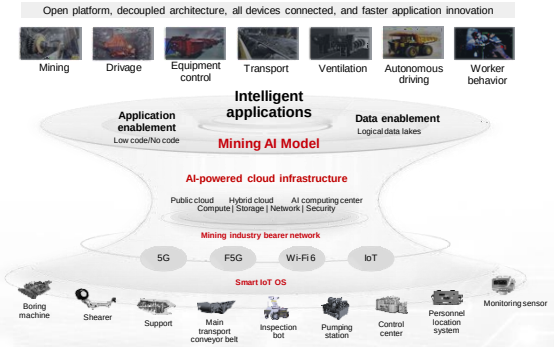
For the fully-mechanized mining face, hundreds of channels of videos are processed in real time. AI stitches the videos together to create a panoramic remote control view, so that coal miners can operate the mine while staying safely above the ground.

Build a smart coal mine based on an industrial Internet architecture

3.0 System solution

Powered by an industrial Internet architecture and a data-AI convergence platform, one large AI model supports E2E task scenarios in a smart coal mine, covering mining, drivage, device control, transport, ventilation, and more working processes.

Unified architecture | Unified standards | Unified data specifications



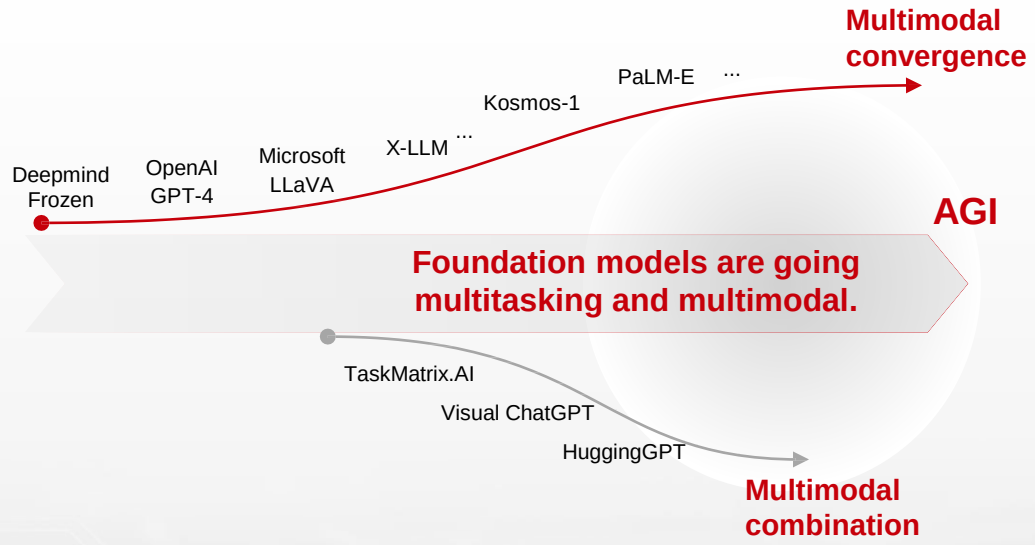
Contents

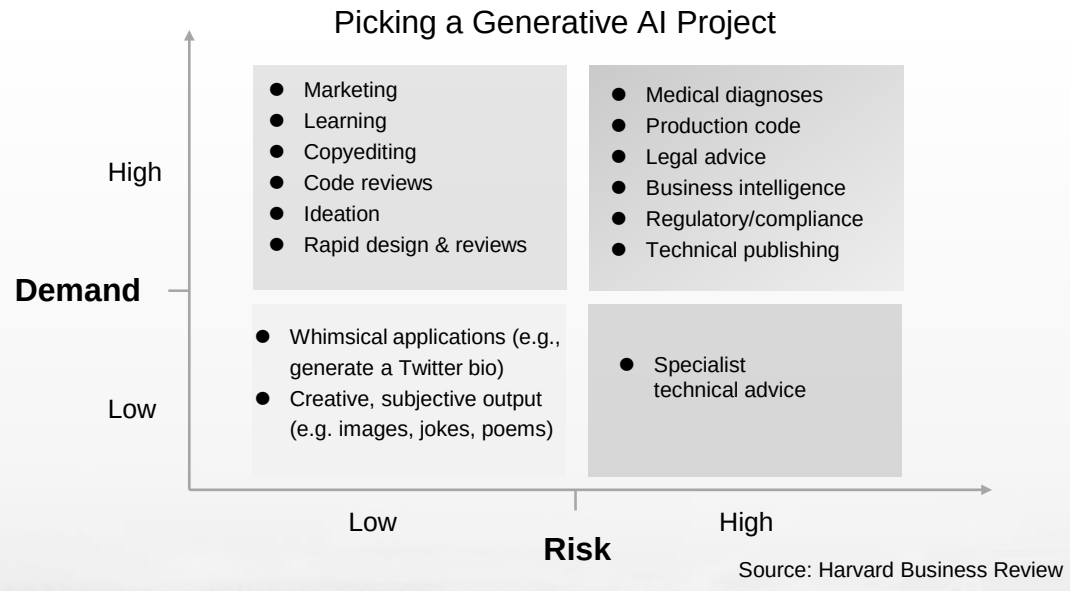
Trend 2: "AI for Industries" Is Accelerating Innovation and Intelligent Upgrades

Powered by foundation models, generative AI has set off a new wave of intelligent upgrades across industries. The key to unleashing the power of foundation models is to adapt them to industry needs. There is huge potential in "AI for Industries" thanks to both the power of foundation models and urgent industry needs.

- **In terms of technology**, foundation models are becoming multitasking and multimodal. Multimodal models offer better generalization by combining the capabilities of CV, LLM, prediction, and more single-modal models. Continuous training on massive datasets continuously improves the generalization capabilities of foundation models, allowing them to be quickly adapted to a diverse range of downstream tasks.
- **In terms of applications**, foundation models are becoming industry-specific to unlock greater value. It is becoming increasingly clear that the future of AI lies in re-training general foundation models on industry-specific, proprietary datasets to generate new, industry-specific foundation models that can be more easily adapted to industry-specific needs.

Today, generative AI applications, such as intelligent Q&A, copywriting, text summarization, machine translation, and computer code generation, are seeing wide and rapid adoption as productivity multipliers. Many industries are rigorously exploring ways to utilize AI and foundation models to redesign or refactor existing business processes. In many cases, the result has been accelerated AI adoption, new business models, and new business growth.





Source: Harvard Business Review

Recommended Course of Action: Use a Decoupled Three-layer Architecture to Build Industry and Scenario-specific AI Models

The industry has long been exploring ways to move from general foundation models to industry-specific models, and a **decoupled, three-layer architecture** has become an increasingly popular choice.

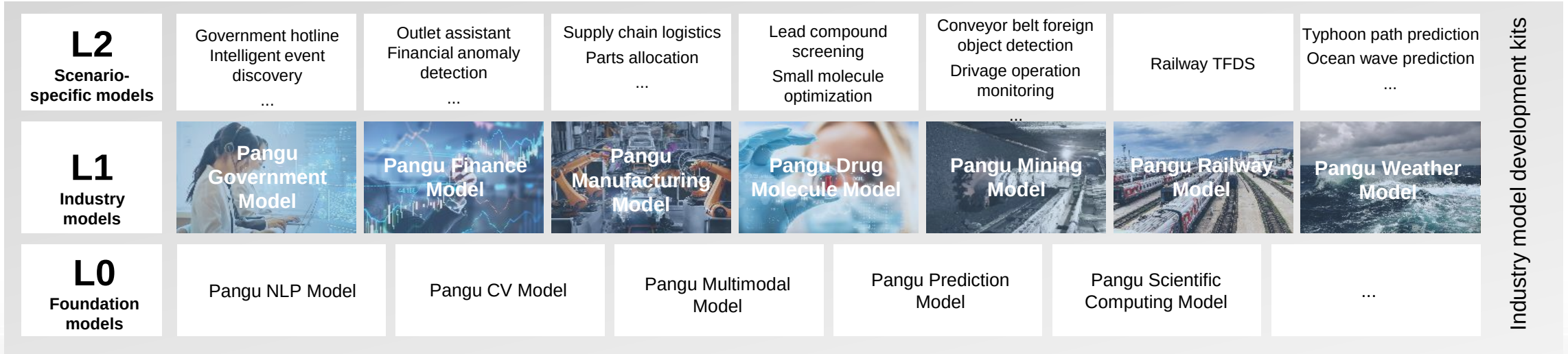
In this three-layer architecture, **the L0 layer consists of a number of general foundation models**, such as NLP, CV, multimodal, prediction, and scientific computing models, which provide general skills to power an endless possibility of industry-specific applications.

- L0 is the pre-training step, where models are pre-trained on huge amounts of general data using self-supervised learning, so the models can acquire massive common-sense knowledge from massive unlabeled data. Huge amounts of training data is needed for this stage, which can be billions of images, massive text, and millions of text-image pairs. The pre-training requires ultra-large compute clusters. Then, we tune the pre-trained models on millions of high-quality examples to let them acquire hundreds of general capabilities. During this process, we use supervised learning and reinforcement learning methods. This way, we turn massive knowledge stored in foundation models into capabilities that can be utilized to solve real-world problems.

The L1 layer consists of industry-specific foundation models that are the result of further training general foundation models on industry-specific datasets. Examples of industry models include those for the government, finance, manufacturing, drug R&D, mining, railway, and meteorology.

- These models have industry knowledge embedded into them, and will offer better performance when applied to industry-specific task scenarios.

The L2 layer provides pre-trained models for specific industry scenarios and tasks that can be quickly deployed off-the-shelf.



Industry Practice 1: Pangu Government Model

Pre-trained on massive amounts of government service data as well as open domain knowledge, such as government service hotline data, government policy documents, and encyclopedias, the Pangu Government Model has developed a range of smart government capabilities, including intelligent Q&A, copywriting, city video sensing, multimodal video understanding, and open-vocabulary object detection, enabling a closed loop of intelligent city event handling that includes perception, understanding, handling, and decision-making.

A unified data platform ingests data from hundreds of thousands of video sources. The Pangu Government Model, tuned on millions of high-quality examples about government service rules and execution, analyzes the video data and discovers events in real time. The Pangu Government Model combines the capabilities of Huawei's Pangu NLP and CV foundation models to enable real-time city event discovery and understanding. For example, after the landfall of a typhoon, the Pangu Government Model can accurately detect fallen trees and shared bicycles, and coordinate a cross-agency response. For example, it dispatches fallen trees to the landscaping and afforestation department, fallen bicycles to the urban administration department, and exposed garbage and waterlogging to the sanitation department. This way, it accelerates emergency responses. City event handling efficiency is improved by over 50%.

Real-time, all-domain sensing	Office collaboration anytime, anywhere	Better services for individuals and businesses	Precise, efficient city governance
Intelligent event discovery	Report generation	Intelligent online services	Government virtual human
Government hotline	Government document writing	Government service assistant	Policy recommendations
...	...	...	...
GOV Governance Digitalization	**GOV Office Digitalization**	**Public Service for Citizen & Business**	**Intelligent Operation Center**

E2E intelligence: perception -> understanding -> handling -> decision-making

Government Q&A bot | **Government copywriting** | **City video sensing** | **Video-based multimodal understanding** | **Open-vocabulary object detection**

Hundreds of billions of parameters

Hybrid network architecture

Pangu Government Model

Pre-trained on massive government service knowledge

Multimodal data understanding and recognition

Industry Practice 2: Pangu Finance Model

Huawei's pre-trained Pangu Finance Model is fine-tuned on bank customers' multi-source, proprietary financial datasets to acquire new capabilities.

The model gives an intelligent assistant to every counter employee working at bank outlets, helping them easily handle all kinds of service requests. By pre-training the model on data and knowledge about banking service procedures, policies, and case studies, the intelligent assistant can automatically generate service workflows and guides for the counter staff based on the customers' service requests. The average number of steps needed to complete a task is reduced from 5 to 1, and the average handling time is reduced by more than 5 minutes.

In the future, we expect to see the Pangu Finance Model deployed for more extensive uses in the finance industry, such as intelligent loan and risk control and analysis.

Intelligent risk control	Intelligent marketing	Smart investment research	Movable property pledge	Automated claim settlement	Intelligent customer service	Intelligent operations
Financial anomaly analysis	Sales script generation	Summary generation	Object recognition	AI + RPA	Agent assistant	Branch assistant
Fraud intention detection	Marketing material generation	Risk evaluation	In-warehouse check	3D damage evaluation	Chatbot	Data analyst assistant
Default risk analysis	Interactive digital human	Viewpoint extraction	Intrusion detection	Intelligent document info extraction	Intelligent outbound calls	Product design assistant

Five financial skills, intelligent assistant for data mining and knowledge management

- Financial policy document-based Q&A
- Cross-modal content understanding and generation
- Multi-task understanding
- Code generation and completion
- Intelligent software interaction and integration

Pangu Finance Model

Flexible deployment mode: public or hybrid cloud

On-demand model extraction

Pre-trained on massive amounts of common-sense knowledge in finance

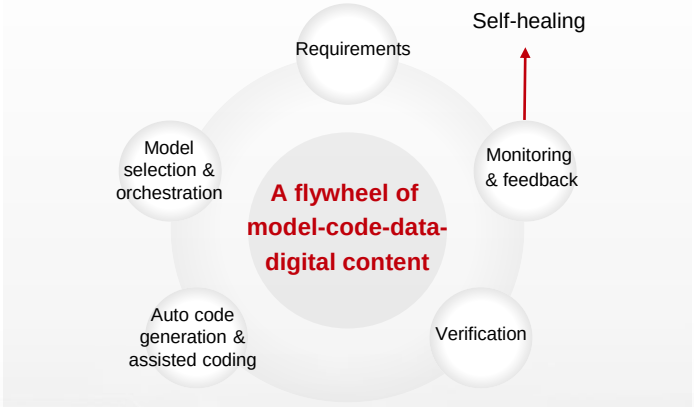
Multidimensional content generation and check

Contents

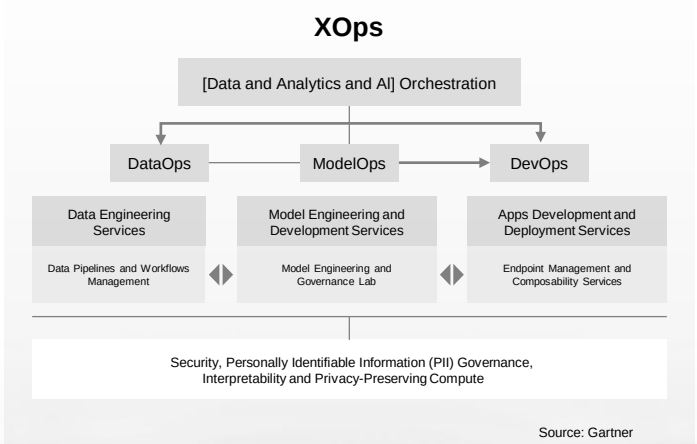
Trend 3: Foundation Models and AIGC Are Transforming the Application Lifecycle from Code-centric to Model-centric

- **From digital to intelligent:** Foundation models facilitate the design and development of acceptance standards, test cases, UI, code, and test scripts for software. Intelligent human-machine interaction helps improve software quality by ensuring that the R&D process is aligned with real user requirements. New intelligent R&D platforms can accurately understand requirements and serve as an intelligent assistant in design, coding, testing, and deployment. An AI flywheel of model-code-data-digital content allows R&D tools to iterate continuously to meet fast evolving requirements.
- **From DevSecOps to multimodal convergence:** The rapid advancement of AIGC has facilitated the seamless integration of data and AI in enterprise applications. These applications will reshape DevSecOps by adopting a model-centric, multimodal pipeline. This pipeline covers not just R&D, security, and operations, but also data, intelligence, and digital content. It is expected to deliver higher productivity than conventional DevSecOps methods.
- **From role-specific to all-in-one:** Many phases in software R&D, such as decision-making, planning, prediction, and coordination, are AI-enabled. AI-powered platforms will augment individual developers' capabilities and also enhance team collaboration and interaction efficiency among different roles. As a result, the structure of R&D teams is shifting from a role-based approach to three distinct categories: product development, architecture, and operations. This helps improve productivity by reducing excessive division of labor in software development.

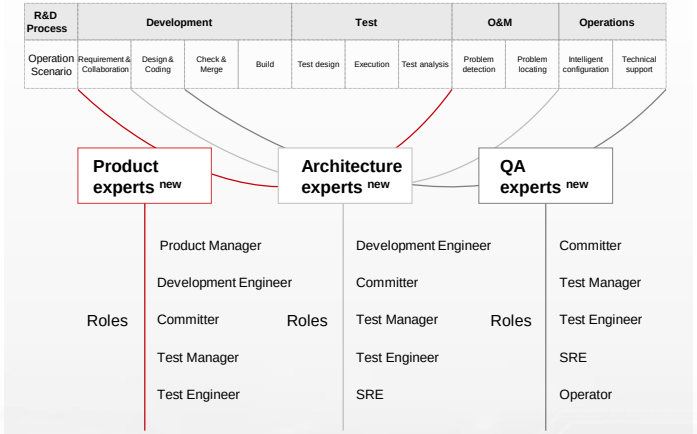
Software 3.0: the era of intelligent software engineering



Multimodal convergence of software R&D workflows



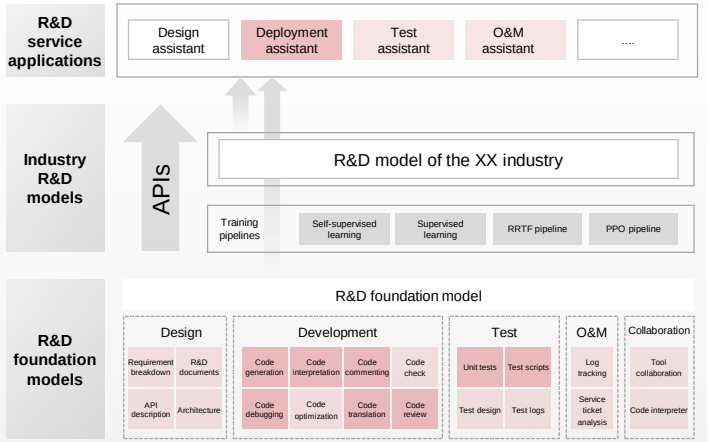
Software development platforms with all capabilities



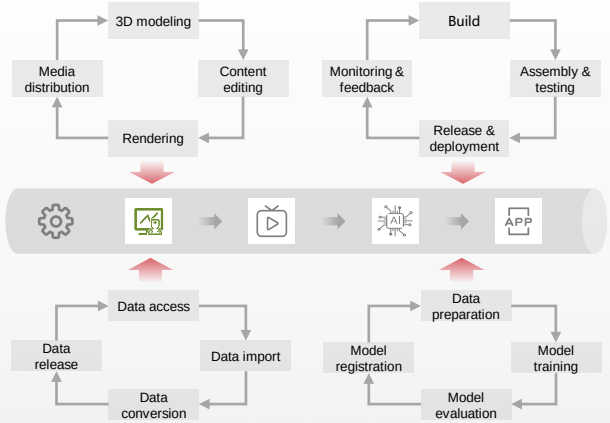
Recommended Course of Action: Use Intelligent R&D Tools and Integrated Open Capabilities to Accelerate AI Development and Ecosystem Building

- **Use intelligent R&D tools:** Provide AI-enabled tools to assist R&D teams in requirement design, development coding, integrated testing, O&M and monitoring, and R&D collaboration.
- **Unify on-cloud development:** Rebuild the software development pipeline with AI, seamlessly integrating data, models, digital content, and software development into a single platform. This pipeline will accelerate the AI application development process across all industries, enabling intelligent transformation.
- **Build a new AI application development framework:** The AI application integration layer is designed for task orchestration, model and application hosting, and application release. On the application development layer, applications can be easily assembled using AI plugins. This layer provides input transformation, AI security, and access to foundation model APIs through the API gateway. Additionally, AI data connectors enable enhanced data search capabilities.

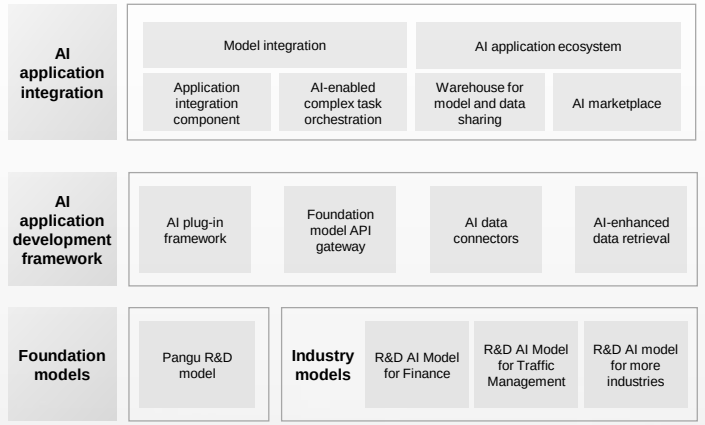
Intelligent R&D tools



Unified on-cloud application convergence pipeline



Multimodal AI application development framework



Contents

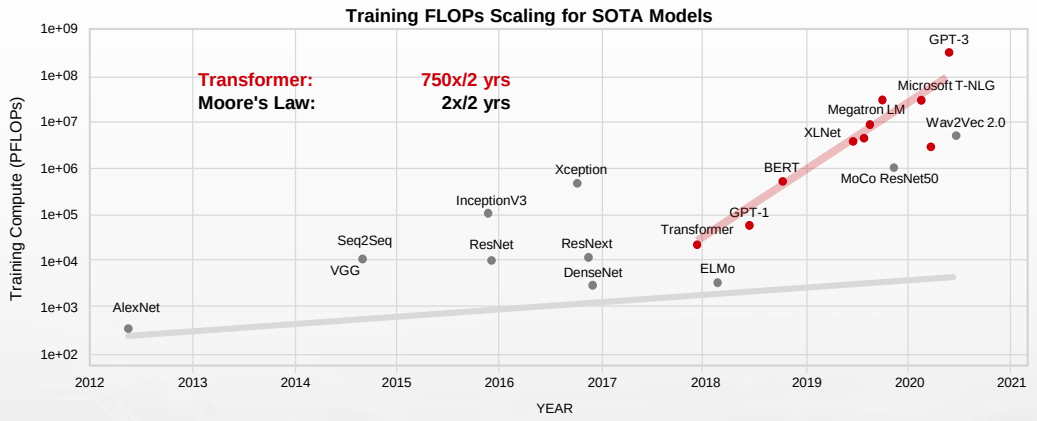
Trend 4: AI Cloud Services Are Becoming the Preferred Way for Enterprises to Build and Power Their Own Large AI Models

AI is reshaping every industry, and foundation models are accelerating this trend. This, along with a technology stack based on AI agents, is showing the world, especially enterprise executives, early sparks of artificial general intelligence (AGI). Leaders in every industry are bracing foundation models wholeheartedly. Some are integrating foundation models into their existing applications using a copilot mode, while some others are restructuring their entire businesses to make the most of foundation models.

- **The scaling laws for AI continue to apply. The sizes of foundation models continue to grow, so do their needs for computing power**. After a large language model (LLM) exceeds a certain critical scale, it begins to show capabilities that its developers did not expect and could not have predicted, because they never occur in smaller models. Such capabilities and features are now commonly referred to as emergent abilities. Developers are likely to further increase the size of foundation models in pursuit of even more powerful AI, and with it comes the growing demand for computing power. In the short term, compute is still mostly used for model training. In the future, only 20% will be used for model training, while 80% will be used for inference. **Reliable, cost-effective AI compute will be crucial**.
- **Foundation models require a new class of compute**. The development and application of foundation models have raised the bar for network bandwidth, GPU capacity, memory bandwidth, and system reliability. Challenges associated with the interconnect bandwidth-compute ratio, memory capacity-compute ratio, and memory bandwidth-compute ratio all need to be addressed. Meanwhile, enterprises are still trying to find their way around foundation models. Their expectations and uses of foundation models and the associated compute needs will keep changing. **Enterprises need efficient, scalable AI compute.**
- **The development and application of foundation models require software & hardware system engineering of unprecedented complexity**. Generative AI as a technology was conceived years ago. Enterprises are now using foundation models and AI in different ways: 1) Integrate AI models into customer-facing applications, and run their own models end-to-end or use third-party models via APIs; 2) offer foundation models for external use by means of proprietary APIs or open-source them; 3) as a cloud service provider or hardware vendor, provide an AI compute infrastructure that supports foundation model training and inference. **Cloud service vendors strengthen their technology stacks and build application ecosystems around foundation models, making them powerful players in the game of AI and foundation models**.

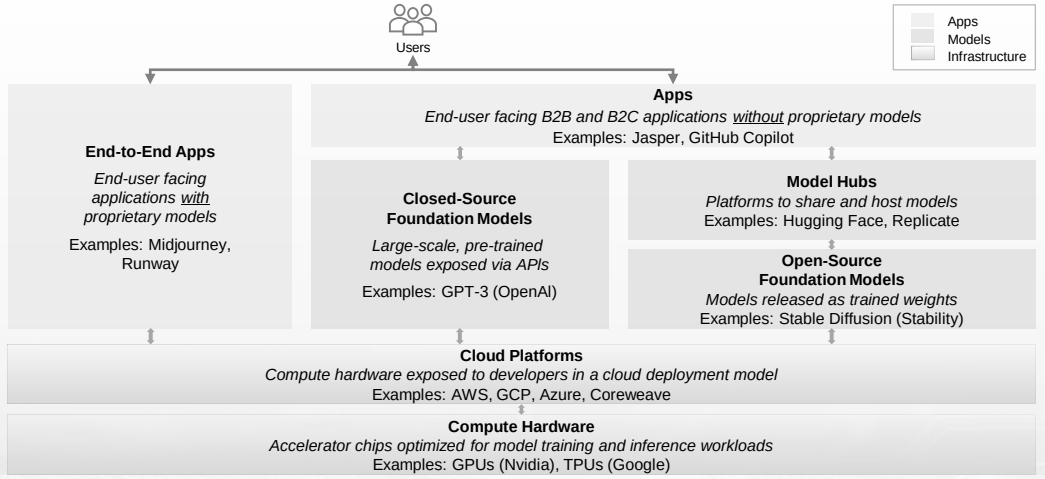
The gap between Moore's Law and the fast growing demand for AI compute is quickly widening

The compute power needed to train a SOTA Transformer model has increased 750-fold in two years.



Source: UC Berkeley RISELab

With the rapid development of foundation model technology ecosystems, cloud vendors have taken their place in the technology stack



Source: https://a16z.com

Recommended Course of Action: Power the Development and Continuous Iteration of Large AI Models with Cloud-based AI Services

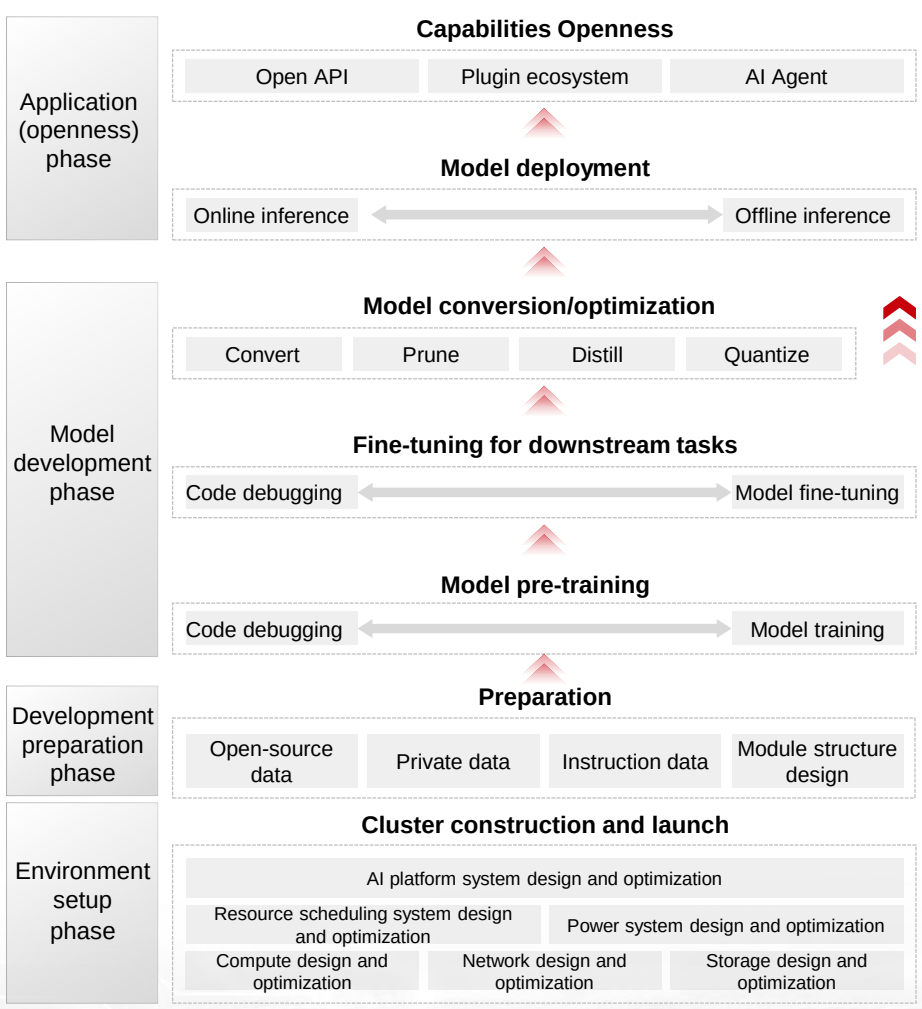
A cloud can provide massive, on-demand, highly scalable compute resources needed for the training and inference of large AI models. Furthermore, developing a large AI model is not just about AI algorithms. Complex system engineering is involved in every step, including data processing, software and hardware optimization, model development, and application innovation. Enterprises can rely on the AI cloud services and AI ecosystems of large cloud vendors to overcome challenges associated with all of these steps. The recommended course of action includes:

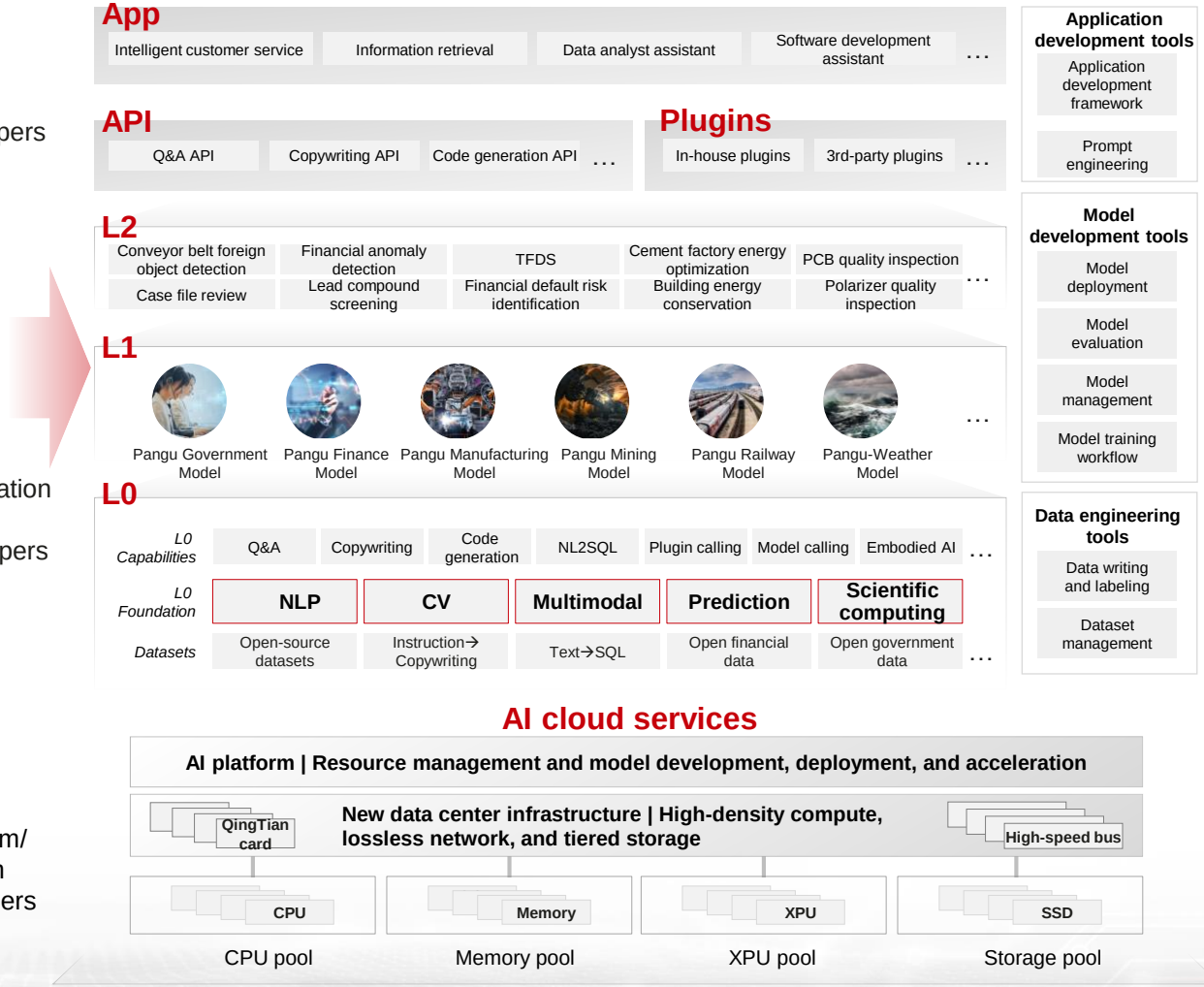
1. **Use the cloud to power AI**: The scaling laws for AI continue to apply. As the sizes of foundation models continue to grow, so do their needs for computing power. The benefits of working with a leading cloud vendor include native software-hardware synergy, end-to-end acceleration, and efficient, scalable AI compute.

2. **Become part of a leading foundation model ecosystem and build industry-focused intelligent systems through joint efforts**: Foundation models are set to reshape every industry. We expect the majority of future AI systems to be powered by the cloud.

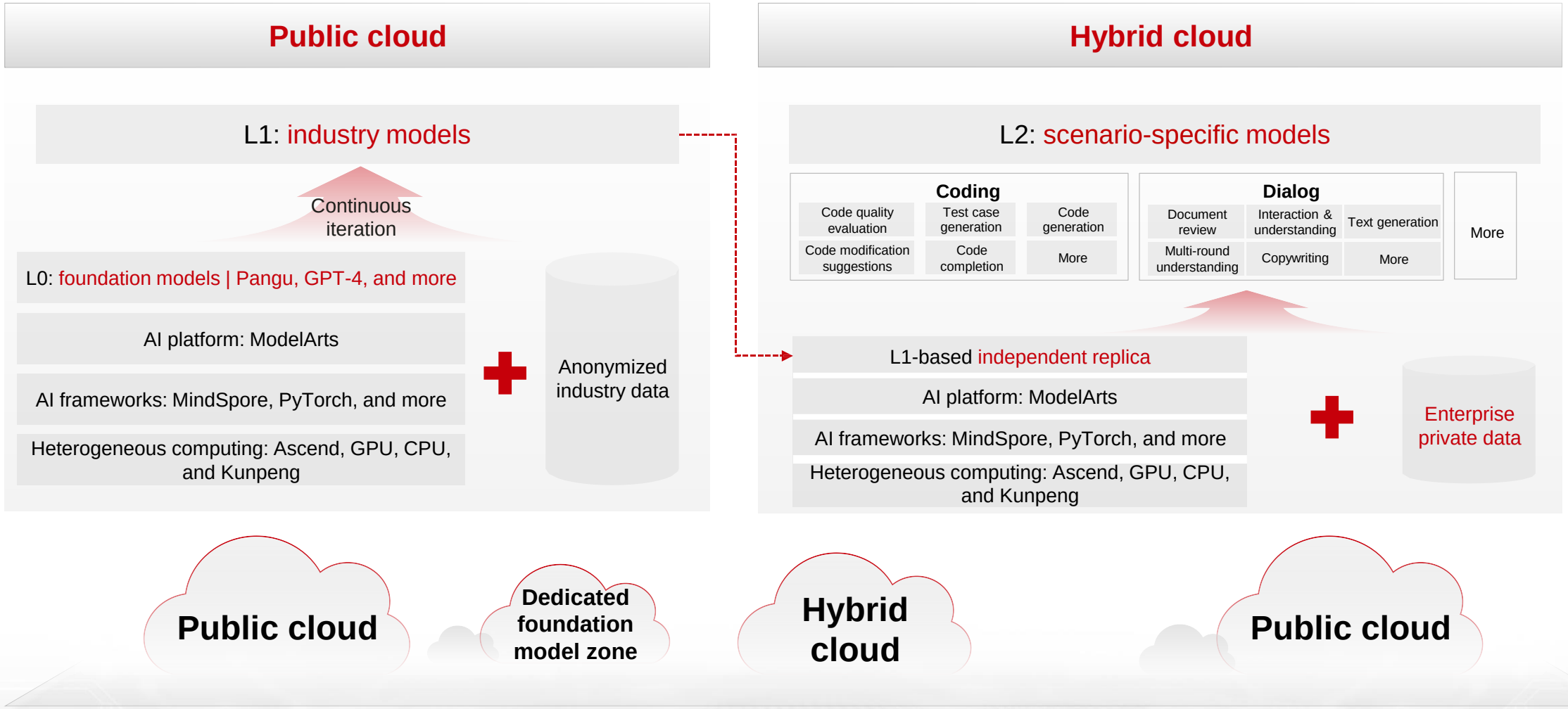
3. **Heterogeneous compute with a multi-cloud strategy**: Business continuity and data security are top concerns for enterprises that are thinking about adopting foundation models. Heterogeneous compute and multi-cloud deployment effectively address these challenges.

Foundation Models Require Software & Hardware System Engineering of Unprecedented Complexity, Which AI Cloud Services Can Help Reduce





Powering Foundation Models with Heterogeneous Computing, and Protecting Core Data Security with a Multi-Cloud Strategy

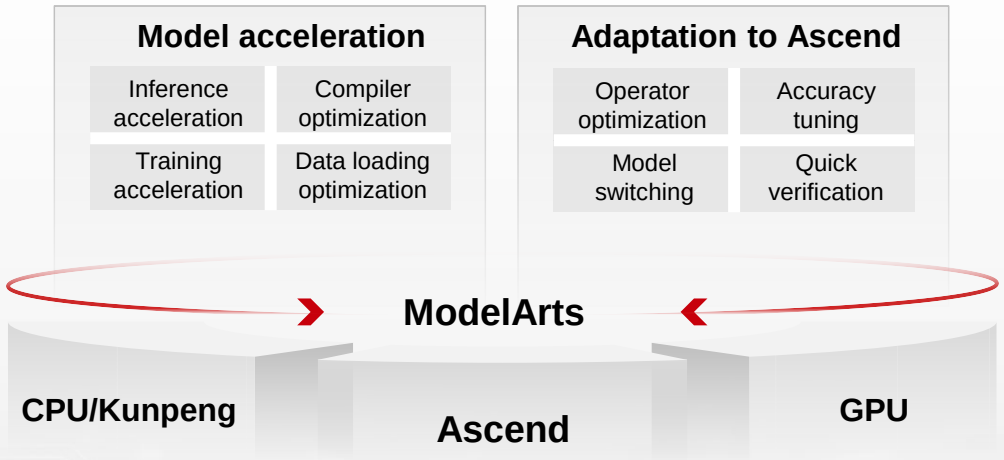


Industry Practice: Powering AIGC and Foundation Models with Heterogeneous Computing to Accelerate Innovation

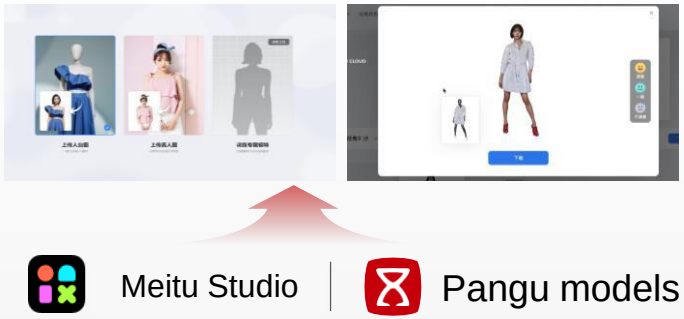
AIGC powered by foundation models has revolutionized digital content creation beyond recognition. Many industries, especially the Internet sector, are actively embracing AIGC. An example of this can be seen in Meitu, which became one of the top 10 apps in 14 countries by launching AIGC functions such as AI painting. In response to explosive growth in compute needs, Meitu obtains massive, efficient heterogeneous AI computing power from Huawei Cloud. Huawei Cloud's Pangu models also power a hugely popular virtual clothes try-on feature recently launched by Meitu. With this feature, Meitu has taken an important step on its journey to becoming an e-commerce service provider.

- **Efficient, cloud-based AI computing power**: Huawei Cloud provides cost-effective heterogeneous AI computing power. The AI platform improves training efficiency by 40% and reduces AIGC model inference latency by more than 33%.
- **Full-stack AI tool chain**: Huawei Cloud ModelArts is compatible with native Kubernetes APIs and mainstream third-party plug-ins for plug-and-play interconnectivity. Huawei Cloud also provides industry-specific foundation model engineering kits, which cover data processing, model training, and application development, accelerating the development of industry-specific foundation models 5-fold.
- **New businesses powered by foundation models**: Meitu blends the capabilities of both open source foundation models and Huawei Cloud Pangu models to build a B2B application design studio. The studio has developed a range of new AI features powered by AIGC.

One-click access to AI cloud services for powerful computing



Business foundation models for B2B innovation



Thank you.

Bring digital to every person, home, and organization for a fully connected, intelligent world.

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI