HUAWEI

Striding Towards the Intelligent World White Paper 2024

Cloud Computing

Thrive with the Cloud:
Reshaping Industries with AI

Building a Fully Connected,
Intelligent World

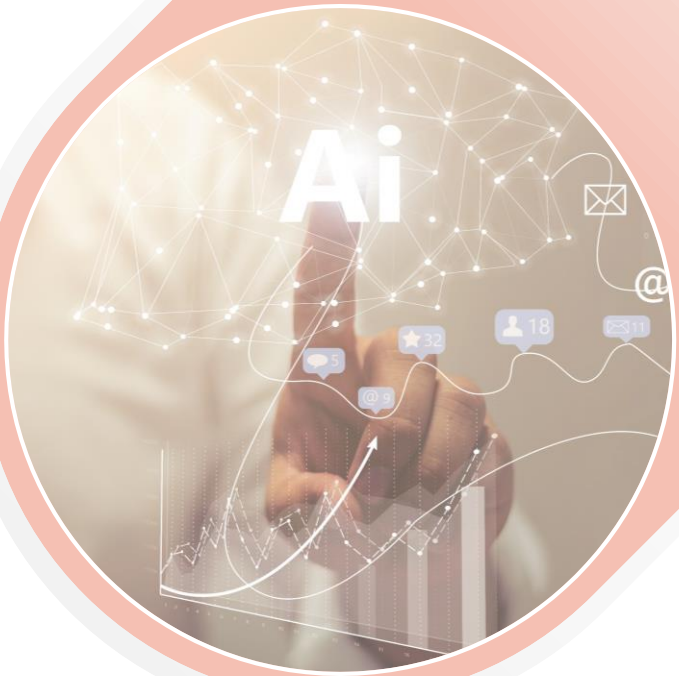
Contents

HUAWEI

01

Trend 1:

AI-native Infrastructure for Elastic, Efficient, Diverse Compute



HUAWEI

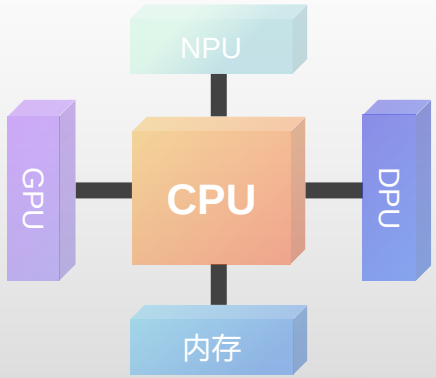
Trend 1: AI-Native Infrastructure for Elastic, Efficient, Diverse Compute

Unbalance in interconnect bandwidth-compute, memory capacity-compute, and memory bandwidth-compute ratios hurts performance, resource utilization, and efficiency. This demands a transformation of the AI infrastructure architecture.

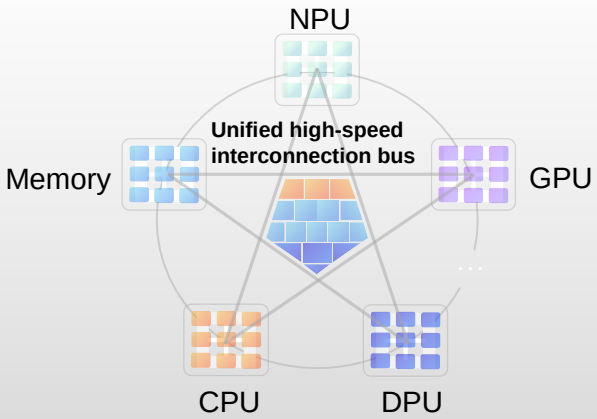
Interconnect bandwidth-compute ratio			Memory capacity-compute ratio			Memory bandwidth-compute ratio		
Transformer	Computational load	↑ 16.8x/2 years	Foundation model	Parameters	↑ 11.2x/2 years	Past 20 years	Hardware compute	↑ 60,000x
	Intranet connection bandwidth	↑ 1.4x/2 years		Memory capacity	↑ 2x/2 years		Memory bandwidth	↑ 100x

AI compute infrastructure: from monolithic to matrix

Primary/Secondary monolithic compute



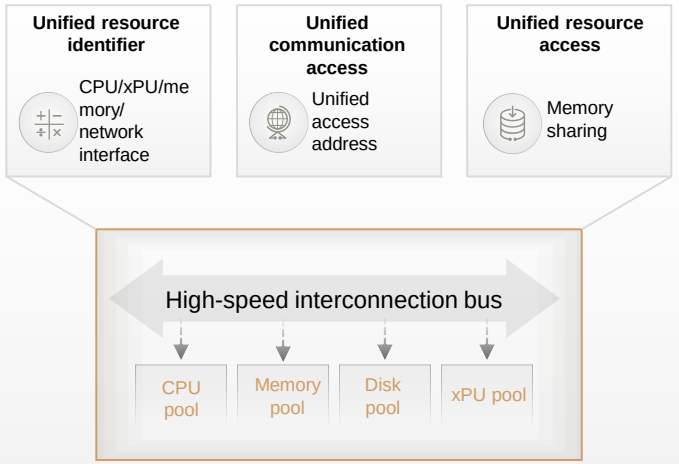
Peer-to-peer pooling matrix compute



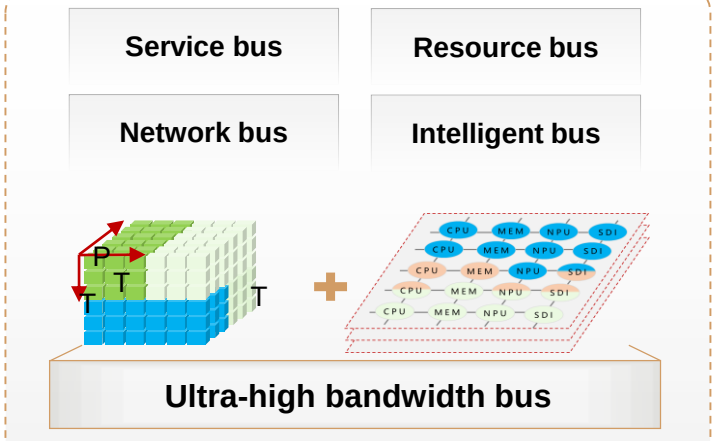
HUAWEI

Recommended Course of Action: Build a Diverse Peer-to-Peer, Pooled, Composable AI Compute Infrastructure

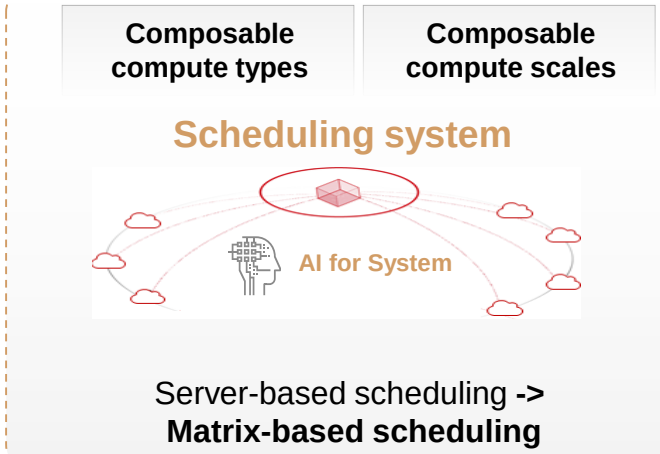
Unified resource pooling / Breaking compute boundaries



Peer-to-peer bus interconnection / Breaking network boundaries



Flexibly composable compute / Breaking usage boundaries



Unified resource pools:
Break compute, storage, and network boundaries for unified virtualization and pooling of diverse resources such as CPUs, NPUs, GPUs, and memory.

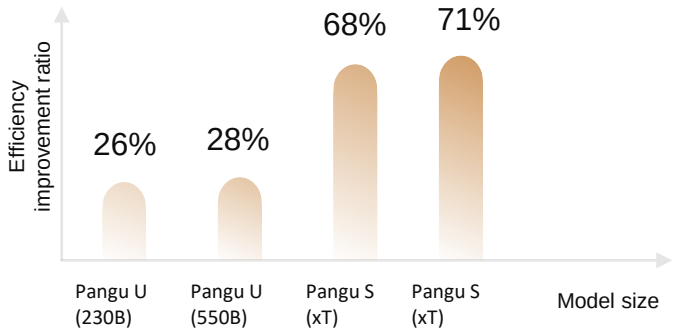
Peer-to-peer bus connections:
Move away from a traditional Ethernet to an ultra-high-bandwidth shared Ethernet bus and break down the cluster linearity bottlenecks through the adaptive topology awareness.

Flexible composition of compute:
Resource allocation is optimized to schedule resources for training of multi-billion parameter or even trillion-parameter models. This unleashes more cost-effective compute and improves asset value.

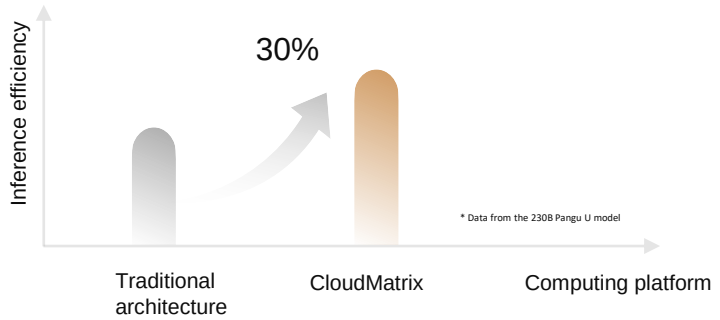
HUAWEI

Industry Practice: An Architectural Upgrade of Huawei Cloud Data Centers Boosts Efficiency in Model Training and Inference

Model training: automatic resource allocation based on model sizes

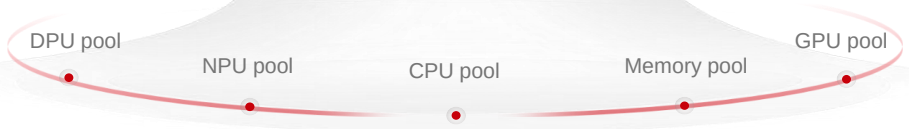


Model inference: higher efficiency through optimal resource composition



CloudMatrix: AI-native cloud infrastructure

Next-generation diverse peer-to-peer full mesh compute



新型高速互联总线

HUAWEI

02

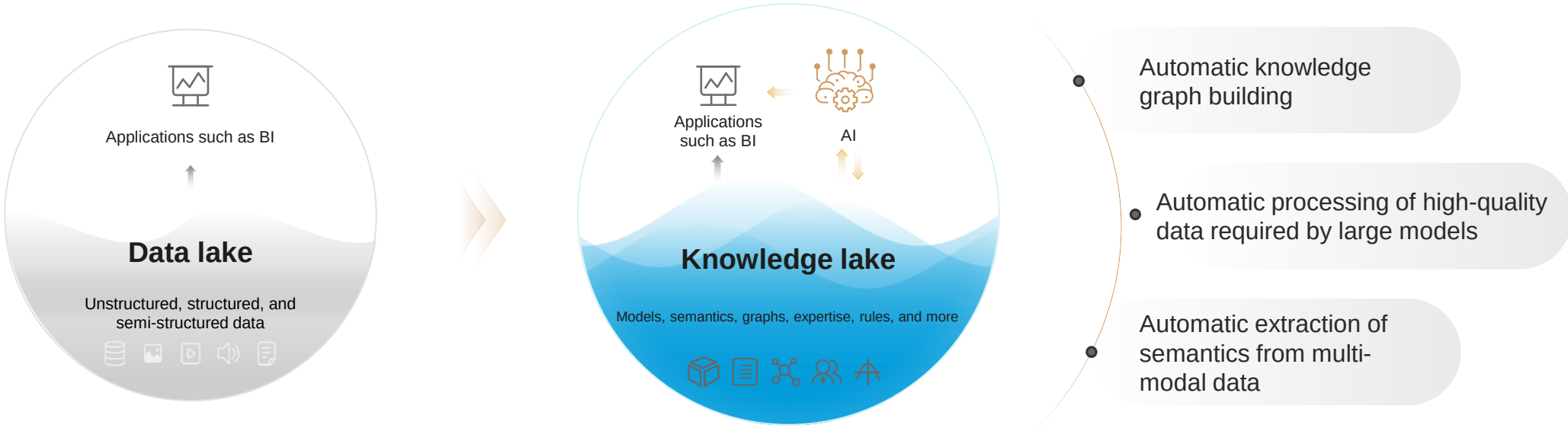
Trend 2:

Knowledge-centric Data Foundations Powering Large AI Models

Trend 2: Knowledge-centric Data Foundations Powering Large AI Models

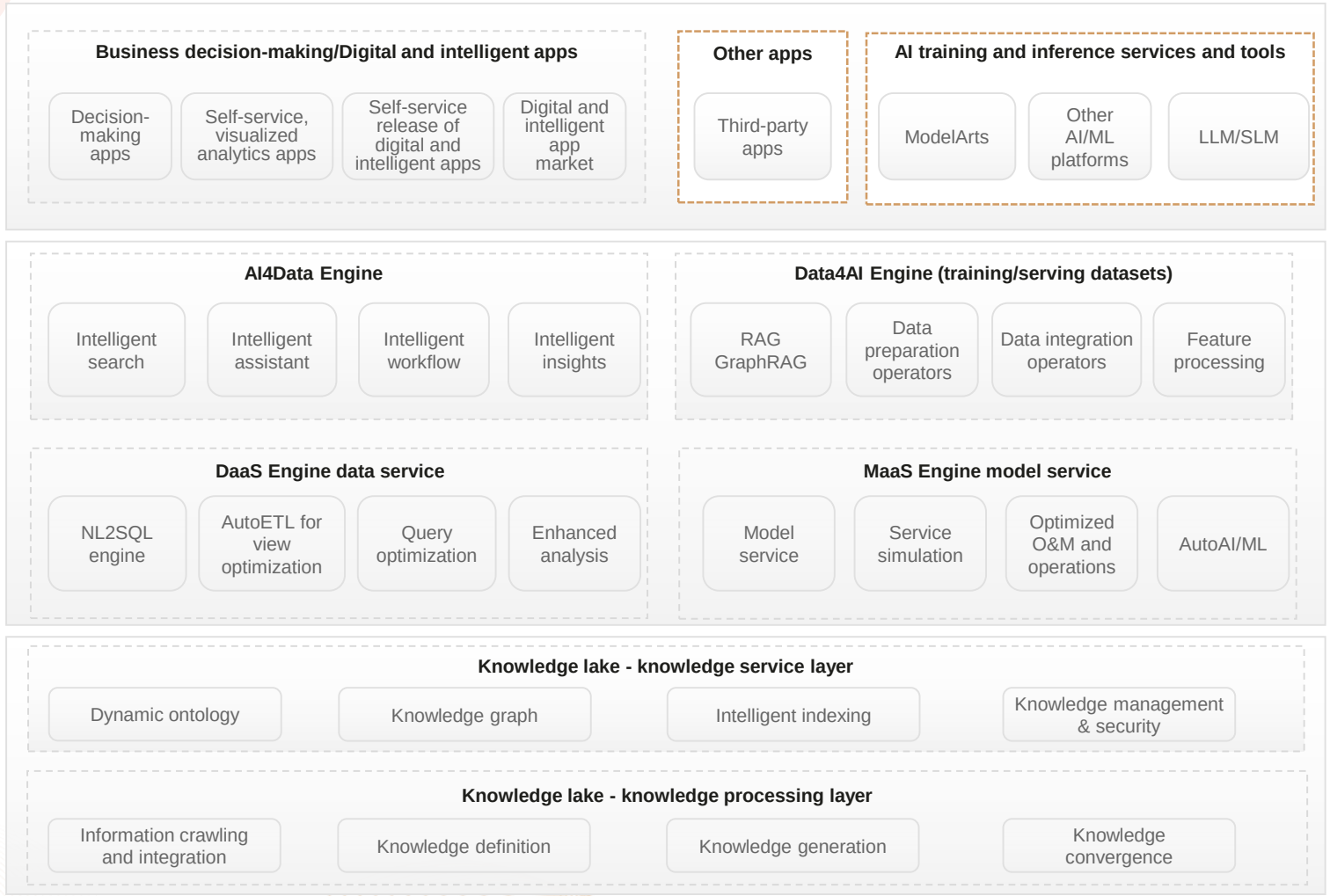
Many companies utilize data warehouses and data lakes for business intelligence. However, these data platforms are not often well-suited for large AI models. To create a knowledge-centric data foundation that can effectively support AI models, companies must clean and extract knowledge from vast amounts of data. The process for building this data foundation includes the following steps:

1. Semantics are extracted from the data in the data lake and associated with business semantics to generate knowledge.
2. AI automatically generates Q&A pairs and image-text pairs, which efficiently provide business knowledge for pre-training and fine-tuning by large models.
3. AI automatically builds knowledge graphs for companies using the data they have obtained. In this way, users in the companies can perform tasks efficiently through natural language dialogs powered by large models.



HUAWEI

Recommended Course of Action: Use a Three-Layer Architecture to Power Large Models



Intelligent decision-making

- **Decision-making apps**: 10x more efficient orchestration of intelligent apps
- **Digital and intelligent app market**: development, release, sharing, and trade of digital and intelligent apps

DaaS/AI services

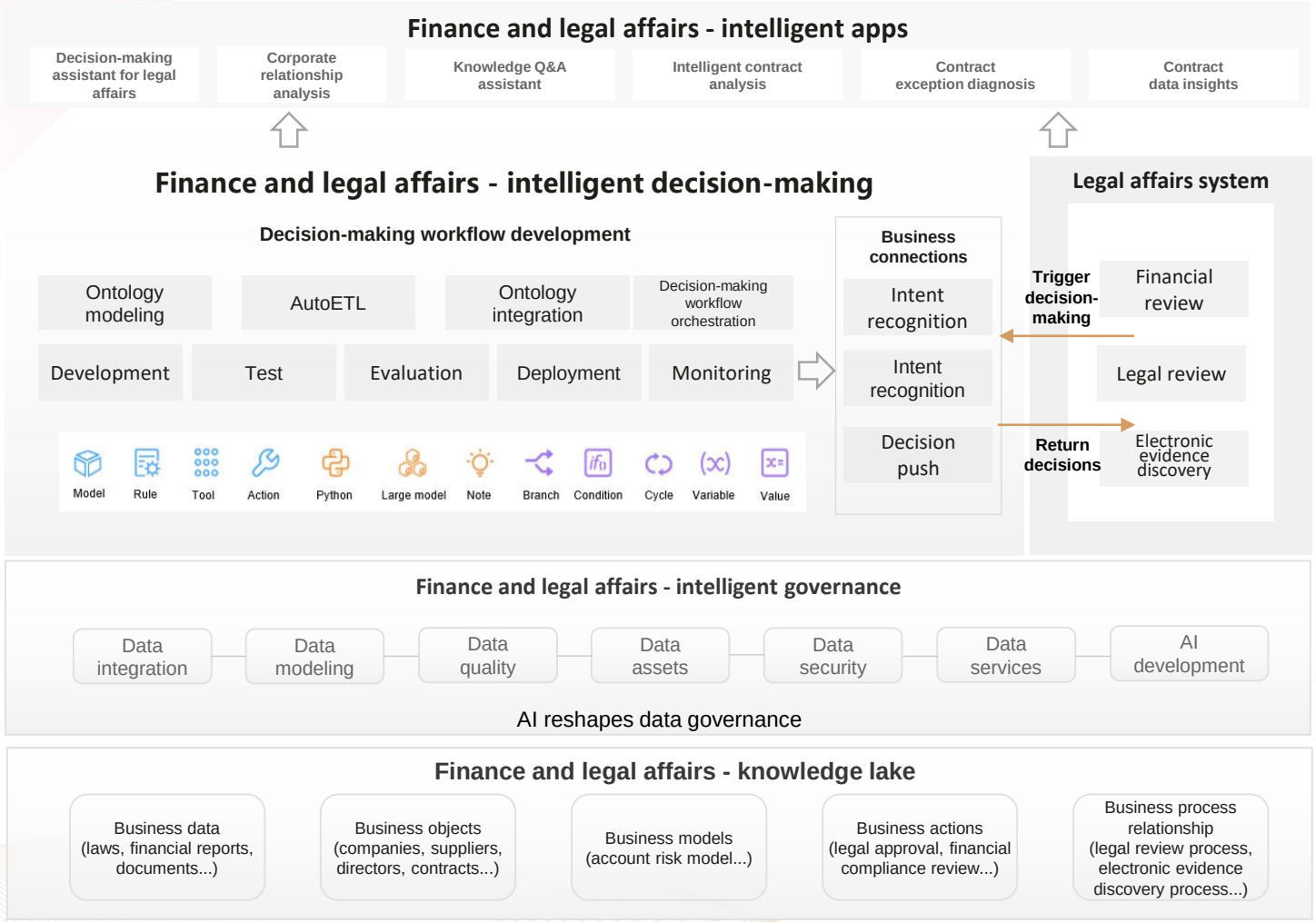
- **DaaS Engine**: unified intelligent data services, including NL2SQL, AutoETL, query optimization, and data insights, with 10x better performance
- **MaaS Engine**: unified model service
- **AI4Data Engine**: AI-powered data governance, analysis, and processing
- **Data4AI Engine**: efficient preparation of high-quality data for training AI models

Knowledge lake

- **Knowledge processing layer**: mines data features in real time, represents them as vectors, generates knowledge graphs based on rules and models, and performs self-learning based on feedback, at 10 times the efficiency of manual operations.
- **Knowledge service layer**: unified knowledge retrieval, knowledge graph, and dynamic ontology knowledge integration

HUAWEI

Industry Practice: Building Intelligent Apps for the Financial and Legal Affairs Departments of a Company



Pain points and solutions

Pain Point	Solution
The IT system cannot perceive and learn the business decision-making logic which is based on expertise.	**Decision-making workflow development: Business users can create decision-making pipelines to integrate expertise into the system.**
Business users struggle to locate and comprehend data, while IT users spend significant time on data development and cleaning.	**1. Business object-business relationship ontology models** **2. Ontology integration: automatic discovery of data related to business systems, establishment of the mapping between physical data and business ontologies, and automatic data integration** **3. AutoETL automatically cleans and transforms data.**
Business users and IT users lack domain knowledge.	**1. With AI, legal affairs users can create decision-making pipelines by themselves.** **2. Intelligent recommendation of simulation algorithm models and pipeline templates**
Personalized decision-making logic cannot be generalized or replicated on a large scale.	**1. Execution of the decision-making workflow and what-if analysis simulation** **2. Automatic extraction of knowledge from the knowledge lake and automatic creation of change pipelines for different scenarios**
Lack of business impact analysis	**1. Natural language Q&A for analyzing the impact of business changes** **2. Intelligent triggering of decision-making APIs for interconnection with peripheral systems**
Lack of business monitoring and unexplainable decision-making process	**1. Visualized decision-making process and logs that can explain the process** **2. Continuous optimization of business operations**

HUAWEI

03

Trend 3:

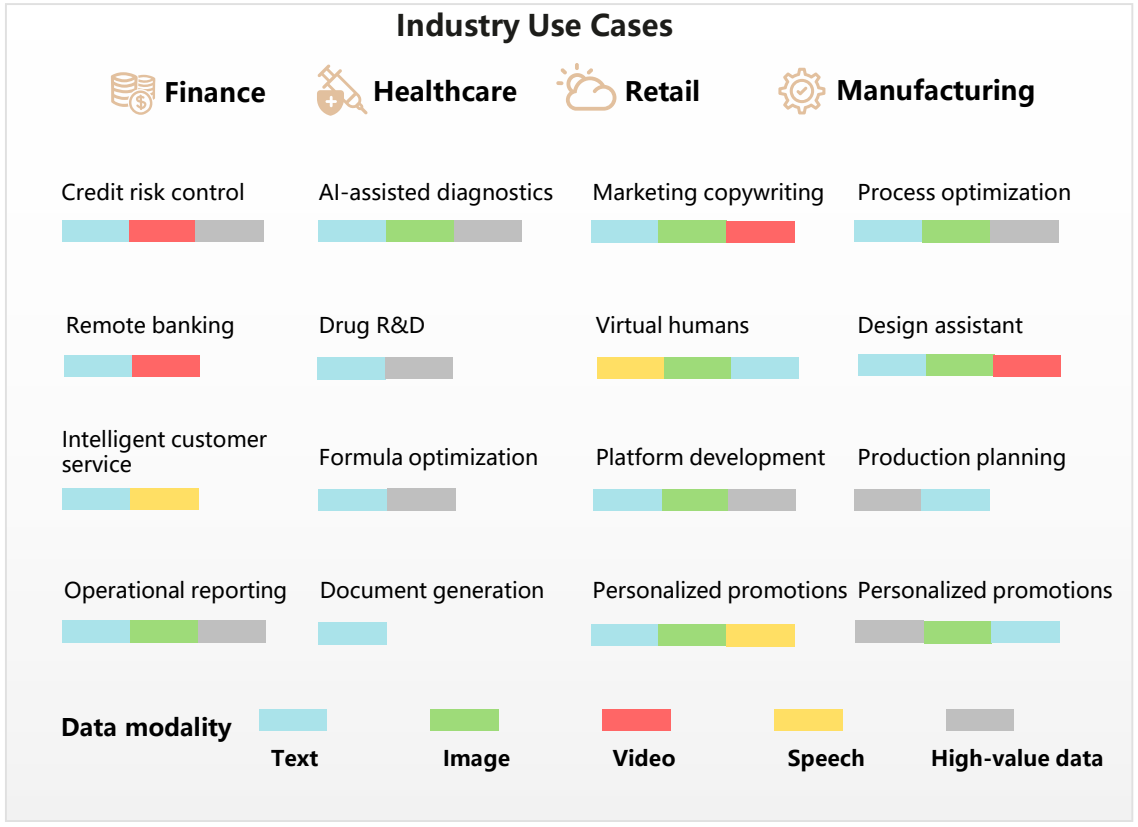
Multimodal, Multi-size Models Meeting Diverse AI Needs

HUAWEI

Trend 3: Multimodal, Multi-size Models Meeting Diverse AI Needs

Today, generative AI is spearheading a new wave of intelligent upgrades for enterprises, with many industries integrating AI into their core production systems. AI must handle various data modalities across sectors, including not just text and images, but also highly specialized and valuable data such as time series data and process parameters in the steelmaking industry, multi-source heterogeneous spatial data in the weather services industry, and medical imaging and gene sequencing data in the healthcare industry. Multimodal models can preprocess these diverse data types and represent them in a unified form. By combining multimodal information, these models enhance accuracy and generalization capabilities, enabling them to tackle more complex problems in specific industrial scenarios.

Industry Use Cases



Key Challenges

Data quality

- A limited supply of high-quality multimodal datasets: Due to high costs in collecting and labeling multimodal data (e.g., videos), there are much fewer multimodal datasets than text-only ones.
- **AI-generated synthetic data addresses the deletion of training data**, : , but challenges lie in generating high-quality synthetic data in a controllable manner.

Algorithm capabilities

- **Higher algorithm and engineering complexity for multimodal models:** representation, alignment, inference, generation, migration, quantization, etc.
- **Modal fusion and collaboration:** Challenges lie in integrating data of different modalities while maintaining uniqueness and complementarity.

Compute capacity

- **Multimodal models are more compute-intensive than single-modal ones.** Typically, given the same amount of information, text < images < videos in terms of the data size. This means multimodal models need to process larger amounts of data.

Trustworthiness and security

- **Explainability and transparency:** Multimodal models are often complex and have even poorer explainability than single-modal models.
- **Security and governance:** Ensuring the responsible use of multimodal AI is crucial.

Recommended Course of Action: Enhance Multimodal Industry Datasets, and Concentrate on High-Value Use Cases

Key Insights	Suggestions
Multimodality enhances human interaction with AI, enabling more creative and higher-value generative AI applications.	1. Find enterprise use cases where multimodal AI can create significantly greater value than single-modal AI. 2. Think about how to use creative methods enabled by multimodal AI to drive productivity.
Multimodal AI faces challenges related to data quality and specialization, while enterprises must address the compute gap.	1. Fully understand and evaluate the technical complexity of multimodal models, and properly integrate input and output data of different modalities. 2. Ensure data quality through data engineering and governance, identify and lower risks in data privacy and security. 3. Find reliable and cost-effective compute resources.
The operationalization of multimodal AI involves a multi-phase process, and the best paths forward are still undetermined.	1. Start from error-tolerant yet high-value use cases to test-run existing multimodal models, thus validating technical feasibility and value creation. 2. Stay open to different technology paths and be ready to adopt the most suitable technology to address the challenges at hand.

Industry Practice: Automated Inspection of High-Speed Trains with Multimodal AI

What most people don't realize is that everyday, every high-speed train must return to a workshop for safety inspection. In China, over 30,000 cars of high-speed trains need to be inspected daily. In the past, large armies of workers performed this task manually in the early morning, which was both time-consuming and labor-intensive.

The Pangu High-speed Railway Model, pre-trained and fine-tuned on extensive datasets specific to high-speed rail, automates a crucial part of this task. It covers all inspection points and achieves over 90% accuracy in fault and defect detection. Using multimodal convergent diagnosis with 2D images, 3D point clouds, and laser spectrum analyzers, the model can accurately identify various complex fault conditions, such as overload, foreign objects, deformation, missing parts, damage, fractures, oil leakage, and loosened parts, with an accuracy exceeding 98%. Additionally, a high-speed rail defect case generation algorithm generates samples for rare faults, enriching existing datasets. Through continuous learning from these enriched datasets, the model can further push fault detection accuracy to over 99%.

Segment Anything and Detect Anything

Fast model fine-tuning
Fault detection accuracy:
90%+

Multimodal convergent diagnosis

2D images + 3D point clouds + laser spectrum analyzer, fault detection accuracy:
98%+

Training sample generation algorithm

Samples for rare faults on high-speed rail can be generated by AI and used to train the model continuously, pushing fault detection accuracy to
99%+

Industry Practice: Controllable Generation of Training Data for Autonomous Driving Models with Multimodal AI

The Pangu Automotive Model employs a technique called Spatio Temporal Controllable Generation (STCG), along with scene-specific video generation, 4D BEV video generation, autonomous driving simulation libraries, and comprehensive road network information, to generate realistic driving videos at scale. Flexible control terms can be added to generate driving videos for various road, lighting, and weather conditions. These help accelerate the commercial readiness of autonomous driving technology.

Adding and deleting traffic participants

Generate empty streets without vehicles.

Add one moving vehicle.

Add multiple moving vehicles.

Integrates perception and decision-making data, supporting model training end-to-end.

Smooth, natural transition between the viewpoints of different traffic participants.

04

Trend 4:

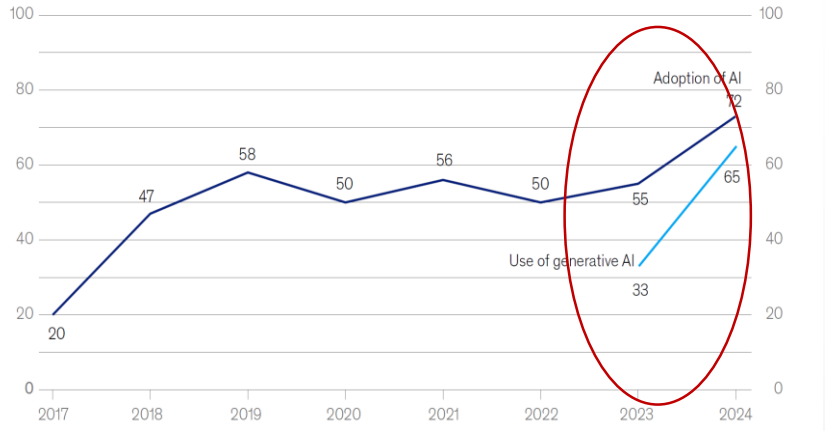
AI Agents, the New Building Blocks of Enterprise AI

HUAWEI

Trend 4: AI Agents, the New Building Blocks of Enterprise AI

Latest survey shows that AI adoption rate has increased from 55% in 2023 to 72% in 2024.

Organizations that have adopted AI in at least 1 business function, % of respondents



Source: McKinsey & Company

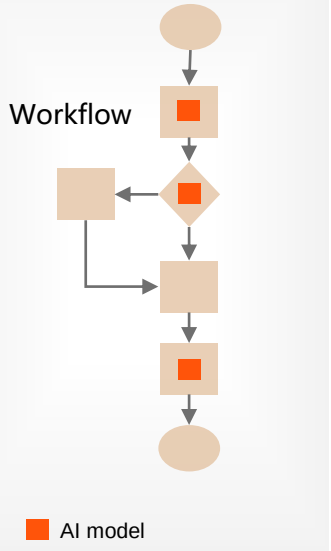
Three trends emerge as foundation model capabilities and AI agents mature:

1.Large models will increasingly replace traditional algorithms in individual tasks like CV and prediction

2.Role-based copilots will collaborate with employees in many functional domains, regardless of industry.

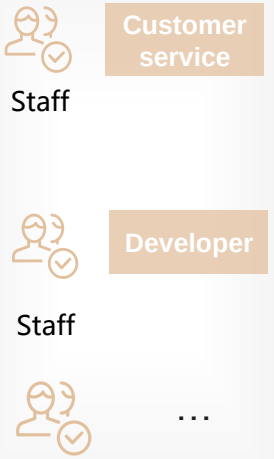
3.Enterprises start to deploy scenario-specific AI agents to handle more complex tasks.

① Embedded AI



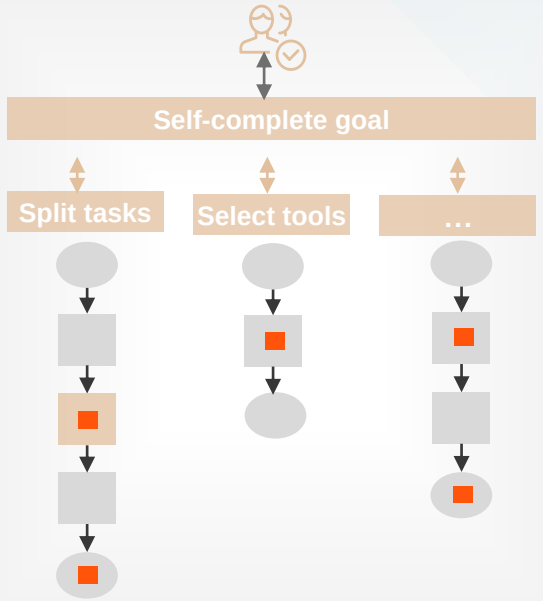
Replaces traditional algorithms with large AI models for better adaptability

② Copilots



Uses role-based copilot applications or AI-powered tools

③ AI agents



Completes human-defined goals autonomously for different scenarios

HUAWEI

Recommended Course of Action: Deploy AI Agents Faster with an Agent Development Suite

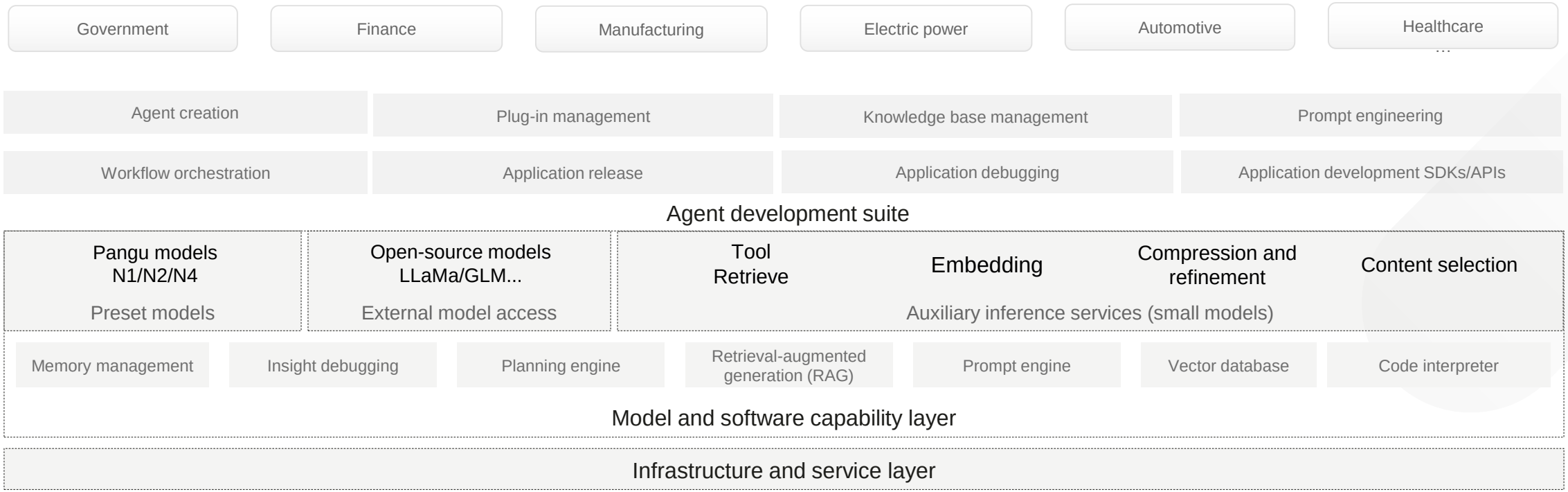
Key steps when developing and deploying AI agents

1，Clarify expectations – problems to be solved, efficiency improvement targets, and expected business value.,

2，Choose an appropriate development suite for flexible deployment, easy integration with existing systems, and powerful AI capabilities.

3，Orchestrate and configure agent applications using the suite‘s plug-ins, knowledge base, and other functions.,

4，Deploy agents gradually – from Q&A agents to more complex customer service/meeting agents. Evaluate and update models after each phase.

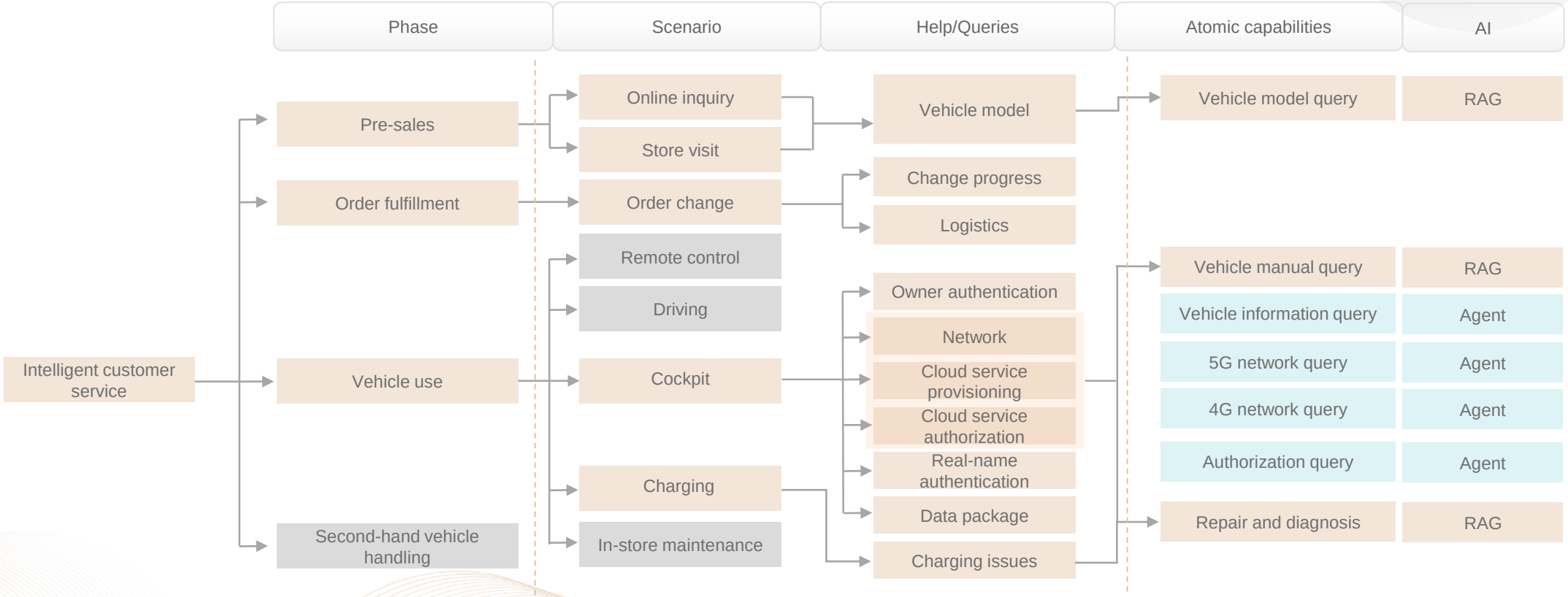


HUAWEI

Industry Practice: Intelligent Customer Service for the Automotive Industry

The agent provides search results and answers questions from car owners about the cockpit, configuration, repair, and diagnosis.

1) Response to cockpit and vehicle model queries
2) Automatic answers to configuration, repair, and diagnosis questions



HUAWEI

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

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

HUAWEI