AsiaInfo
亚信科技

White Paper for

AISWare Usights • TAC MaaS V4.1

An enterprise AI platform that streamlines the AI lifecycle from model development to real-world deployment. It bridges the gap between general-purpose LLMs and industry-specific needs, resolves inefficiencies in operations and resource usage, and helps businesses shift toward AI Native services.

Disclaimer Statement

AsiaInfo Technologies (China), Ltd., hereinafter "AsiaInfo Technologies" or "AsiaInfo", exclusively owns all intellectual property rights, including but not limited to copyrights, trademarks, and patents, as well as technical secrets related to the Product and its derivatives, along with all related documentation, including all information within this document and any attachments.

The information within this document is confidential and intended solely for use by the recipient(s) designated by the user. Without prior written permission from AsiaInfo Technologies, any user of this document shall not take any actions with respect to this Product or the information contained herein to any third party, including but not limited to managers, employees, and affiliates other than the designated recipients. The aforementioned actions encompass but are not limited to developing, updating, compiling and decompiling, assembling, lending, transferring, selling, disclosing, authorizing, distributing, or any other actions. Nor shall any such third party be permitted to use the Product and the information in this document for any purpose whatsoever.

Without prior written permission from AsiaInfo Technologies, users shall not copy, modify, or distribute this document for any purpose. Altering, removing, or damaging any trademarks used in this document is strictly prohibited.

This document is provided as original, and AsiaInfo Technologies makes no warranties regarding the correctness, accuracy, reliability, or any other aspect of this document or its consequences after use. All information in this document is subject to further modification without prior notice, and AsiaInfo Technologies disclaims any responsibility for errors or inaccuracies that may be present in this document.

AsiaInfo Technologies shall not be held responsible for any and all types of liabilities, infringements, or damages resulting from the use of the Product or the information within this document. This exclusion of liability encompasses all forms of damages, including but not limited to direct, indirect, incidental, special, or punitive damages, regardless of whether AsiaInfo Technologies was notified of the possibility of such damages beforehand. The exclusion of liability applies to all forms of legal claims, including those arising from negligence or other torts.

AsiaInfo Technologies' products may include third-party software. Please refer to the copyright statements in the third-party software documentation for details.

AsiaInfo Technologies Limited (Stock Code: 01675.HK)

AsiaInfo Technologies is a leading software product and service provider in China, boasting extensive experience in software product development and large-scale software engineering implementation. With 30 years of deep market presence, AsiaInfo has advanced technological capabilities and numerous successful cases in 5G, cloud computing, big data, artificial intelligence, the Internet of Things (IoT), smart operations, and business and network support systems. AsiaInfo's clientele spans industries including telecommunications, broadcasting, energy, government, transportation, finance, and postal services.

In 2022, AsiaInfo acquired iResearch Consulting Group Co., Ltd. (iResearch Consulting) and integrated it into the new brand iDigital, expanding AsiaInfo's capabilities from product development, delivery services, data operations, and system integration to consulting planning and intelligent decision-making, establishing itself as a leading provider of end-to-end capabilities in digital intelligence.

AsiaInfo is committed to empowering various industries with technologies such as 5G, AI, and big data, collaboratively creating digital value with customers. AsiaInfo aims to lead in both products and services, focusing on continuous product development in the areas of data and intelligence, cloud and network, IT, and middle office products. The cloud and network products maintain international leadership, while data and intelligence products achieve domestic leadership and some international advancements. In the IT domain, AsiaInfo's products stand at the forefront within the domestic landscape.

In the future, AsiaInfo strives to become the most trusted leader in digital intelligence, leveraging its comprehensive capabilities in the field to innovate customer value and contribute to digital transformation.

Certificates (Part)

Capability Maturity Model Integration (CMMI) Certificate L5

Cloud Managed Services Capability Assessment Certificate: Excellent Level

Digital Trusted Services - R&D Digital Governance Capability Certificate

Enterprise Credit Grade (AAA) Certificate

Information System Construction and Service Capability Assessment CS L4

ISO9001 Quality Management System Certificate

ISO20000 IT Service Management System Certificate

ISO27001 Information Security Management System Certificate

Service Certificate of Information System Security Development L2

Service Certificate of Information System Security Integration L2

Awards (Part)

IDC Future Operations Leadership

Leading AI Enterprise in China

Leading Enterprise of Advanced Smart City

The Most Innovative Enterprise Award of China Digital & Software Services

The Most Influential Enterprise in China Software Industry

The Most Influential Industry Brand in China Software & IT Services Sector

The Most Valuable Brand in China Software & IT Services Sector

Top 100 China Competitive Software & IT Services Enterprises for consecutive years

Top 100 China Software Business Revenue List for consecutive years

Top 50 Socially Responsible Enterprises in China Electronic Information Industry

Contents

1 Executive Summary

General-purpose LLM pushes AI into a new stage in 2018. Generative AI has constructed a fundamental architecture for AI and driven the evolution of content generation approach for more industry innovation and diverse human-computer interaction. Customers are now diving in to deeper integration of LLM and core businesses for sustainable growth.

AISWare Usights • TAC MaaS is an end-to-end AI platform serving as a unified MaaS foundation for enterprises. It supports the full lifecycle of LLM/SLM from R&D to implementation, bridging between general LLM and industry applications and driving enterprise's digital intelligence progress with a cost-effective approach.

AsiaInfo is developing AI Native product system for more competitive product capabilities and better customer services. As the foundation for AI Native products, AISWare Usights • TAC MaaS carries all basic functions and supports distributed secondary training, fine-tuning, and adaptive capabilities for LLMs. It can optimize the workflow of ModelOps and FMOps with three augmented integration capabilities and featured functions.

This white paper will elaborate on AISWare Usights • TAC MaaS through the following sections, including product overview, product portfolio, basic functions, featured functions, unique advantages, scenario-based solutions, customer success stories, as well as certificates and awards.

2 Abbreviations and Terms

Terms explanation for AISWare Usights • TAC MaaS are shown below:

Table 2-1 Term Explanation

Abbreviation or Term	Full name	Explanation
AI	Artificial Intelligence	The capability of computer systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making.
AI Native	AI Native	Technology systems where AI is embedded as a core, intrinsic component throughout the entire lifecycle of operations, functions, and processes.
API	Application Programming Interface	A set of rules or protocols that enables software applications to communicate with each other to exchange data, features and functionality.
Cognitive Augmentation	Cognitive Augmentation	The use of cognitive computing systems that supplement human thinking and analysis, providing decision support and automation of routine tasks.
FMOps	Foundation Model Operations	The comprehensive set of practices, processes, and frameworks used to manage the entire lifecycle of foundation models from

Abbreviation or Term	Full name	Explanation
		deployment to ongoing optimization and monitoring.
LLM	Large Language Model	A category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content.
MaaS	Model as a Service	Host pre-trained models on cloud infrastructure and making them accessible via APIs.
ModelOps	Model Operations	A comprehensive framework that focuses on the effective management and operations of AI models.
Prompt Engineering	Prompt Engineering	The process of writing, refining and optimizing inputs to encourage generative AI systems to create specific, high-quality outputs.
RAG	Retrieval-Augmented Generation	The process of referencing an authoritative knowledge base outside of a LLM's training data sources before generating a response to optimize its output.
SLM	Small Language Model	AI models smaller in scale and scope than LLMs and capable of processing,

Abbreviation or Term	Full name	Explanation
		understanding and generating natural language content.

3 Product Overview

With the lightweight and composable architecture, AISWare Usights • TAC MaaS ("TAC MaaS" or the Platform) integrates AI computing power scheduling and full-stack capabilities to serve enterprises as an AI platform that bridges the silos between the LLMs and industry applications. It solves problems such as low resource utilization and isolated O&M and empowers intelligence upgrades from process-driven to cognition-driven.

3.1 Trends and Challenges

LLM, advancing with Generative AI, is now revolutionizing the society structure and cognitive paradigm and serving as the core for intelligence upgrades across industries. Vertical industries, such as telecom and finance, are exploring the commercialization and implementation of LLM as the AI infrastructure in various scenarios, such as customer services, business intelligence, and network optimization.

According to Gartner's research, AI engineering has reached the peak of market expectations, making it a critical technological trend. Reports from CAICT and Tencent suggest that industry-specific LLMs are playing an increasingly important role in driving digital upgrades, with Model-as-a-Service ("MaaS") as the core for full-scale deployment and implementation.

But there still exist bottlenecks in data collection, high training costs, and lack of standards, as well as security risks. In the future, the LLM will evolve towards large-scale training, strong domain integration, and efficient LLM+SLM collaboration for better decision-making. A MaaS-based platform will bridge the technology and applications for enterprises.

3.2 Product Positioning

TAC MaaS provides out-of-box MaaS services and empowers efficient LLM application in business scenarios with smarter capabilities for GPT toolkit.

It accumulates AsiaInfo's rich experiences in services and businesses and data/algorithms to build an AI middle office with high security, stability, and

efficiency for customers in telecom, energy, and transport. It advances AI democracy and engineering to empower industry intelligence transformation.

4 Product Portfolio

TAC MaaS provides end-to-end AI technologies around the full process from data preparation, model fine-tuning/management/deployment, service development, to application. With open AI services, it empowers up-layer applications to transform the enterprise ecosystem.
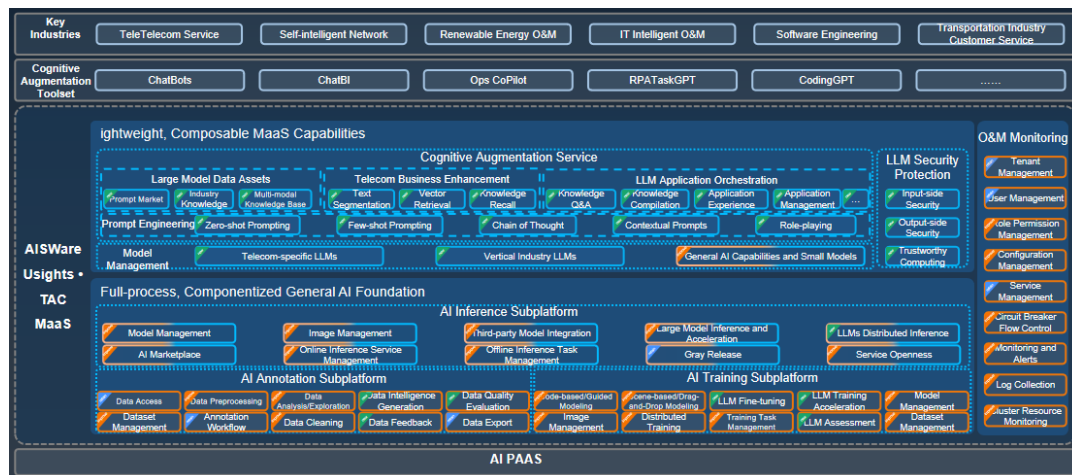

Figure 4-1 Product Portfolio

- **End-to-end general component-based AI foundation**

TAC MaaS visualizes all development capabilities from data preparation to AI service build-up, and integrates/schedules basic physical resources, computing power, and development environments. It satisfies all scenario-based requirements.

➤ **AI annotation platform**

It annotates diversified scenarios from text, images, and audios/videos. It can automatically pre-process massive unstructured data to enhance model quality and reduce the manual annotation costs. It also supports automatic LLM-based annotation.

➤ **AI training platform**

It provides a microservice-based model development environment for LLM/SLM training and evaluation. By decomposing the training engine and services, it enables independent environments for each development tasks and users with customization to solve resource conflicts.

> ➤ **AI reasoning platform**

After training the model, users can build online/offline reasoning capabilities from mirror setup/activation to registration within this unified platform.

• **Lightweight and composable MaaS**

TAC MaaS leverages visualization tools such as LLM/SLM management, prompt engineering, and knowledge bases to augment cognition for knowledge integration and agent development, allowing LLM commercialization and industry empowerment.

> ➤ **Model management**

Based on AI reasoning platform, it hosts and pre-integrates LLMs from AsiaInfo, customers, third parties with domain-specific SLMs with a standard and unified API.

> ➤ **Augmented cognition**

TAC MaaS supports rich industry-specific prompt templates and LLM augmented cognition. With RAG-based multimodal knowledge integration and edit, it helps enterprises to develop LLM applications in an efficient way.

> ➤ **LLM guardian**

Focus on LLM security from the perspectives of input/output security and trusted computing.

• **O&M monitoring**

Support multi-dimension O&M for platforms, tenants, and projects, including tenant/user/role management, project support, log collection, and cluster resource monitoring.

5 Basic Functions

Basic functions are shown as below:

Table 5-1 Basic Function List

Module	Description
Dataset Management	Access and pre-process data to provide model training with high-quality sample dataset: • Manage dataset for model development/reasoning • Manage external multi-source data • Pre-process data such as image and text
Data Annotation	Annotate image, text, and audio/video: • Manage annotation task and tools • Preset various data annotation scenarios for different demands and allow users to develop easy-to-use UI • Rich annotation components for flexible annotation scenario template customization • LLM-based AI annotation • Collaborative annotation workflows, allowing task decomposition
Model Development	Support diversified automatic model training modes: • AI-powered scenario-based modeling for users • Automatic and visualized wizard-based modeling for experts • Drag-and-drop modeling for service experts without any coding requirements • Coding-based modeling for algorithm experts • Allow users to publish codes as wizard-based training templates

Model Hosting	Host self-developed/third-party (open-source and COTS) models: • Model versions • Third-party models • Model conversion accelerator and edge adaption
Model Deployment	Model reasoning: • Online SLM/LLM reasoning task management • Offline reasoning task management • Gray-scale task release supports A/B testing when upgrading online reasoning services, and provide different model versions to correspond to the same service.
Mirror Management	Training mirror management for model development, supporting online customization • Reasoning mirror management during deploying and reasoning, supporting LLM reasoning mirror API control
Open Service	Open AI capabilities by API: • API gateway to unify the access portal • Service permission control • API User management
Operations	Manage platform, tenants, and projects for system operations: • User/role management • Organization management across tenants and projects • User Center to manage messages and change passwords • System management for customized configuration
Container Management	Manage and monitor resources of the container platform

	• Computer memory resource management
	• Host O&M monitoring
	• Container resource scheduling

6 Featured Functions

6.1 LLMs/SLMs Hosting and Pre-integration

TAC MaaS hosts and pre-integrates LLMs and SLMs from AsiaInfo, third parties, and customers themselves. By combining industry experiences and expert knowledge, it provides rich AI capabilities, industry-specific models, and out-of-box industry templates, supporting customers to quickly train data, upgrade and deploy reasoning services.

Figure 6-1 Pre-integrated LLM/SLM Capabilities

6.2 Scenario-Based LLM Fine-Tuning

TAC MaaS secures scenario-based LLM fine-tuning lifecycle from data preparation, training/validation, and evaluation.
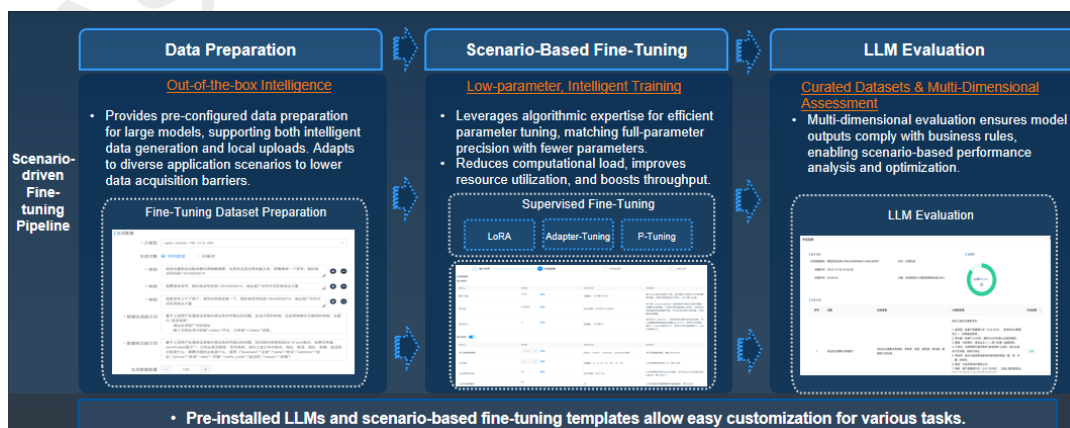

Figure 6-2 Scenario-based LLM Fine-Tuning

6.3 Comprehensive LLM Evaluation Intelligence

TAC MaaS provides automatic tools for quantifying LLM outputs and performance across scenario-based and custom assessments. Based on the open-source framework, it sets up an evaluation workflow to reduce time consumption for frequent parameter/tool iteration.
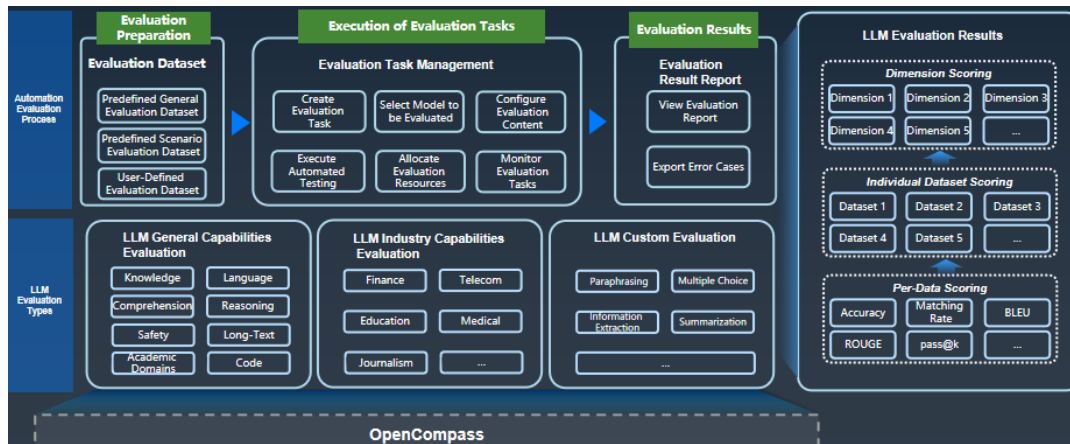

Figure 6-3 Comprehensive LLM Evaluation Intelligence

6.4 International Signaling Standard and Elastic AI

TAC MaaS provides an international AI-enabled signal interaction system based on Open API, as well as standard and programmable gateways. It also supports elastic scaling of AI resources to enable large-scale distributed training and inference with a dual-active architecture.
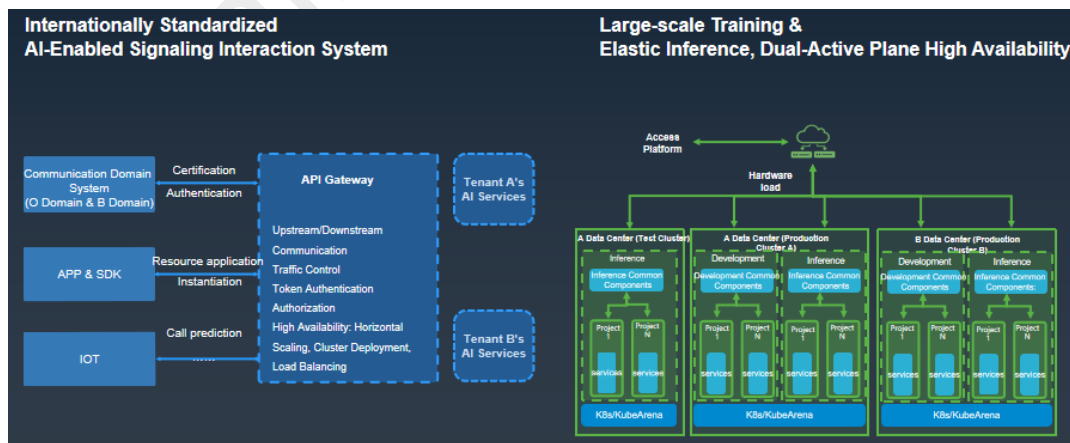

Figure 6-4 International Signaling Standard and AI Elastic Scaling

7 Unique Advantages

7.1 Augmented-Cognition LLM Empowers Industries

TAC MaaS provides unified and open hosting with learning and knowledge migration to augment LLM cognition for industrial applications.
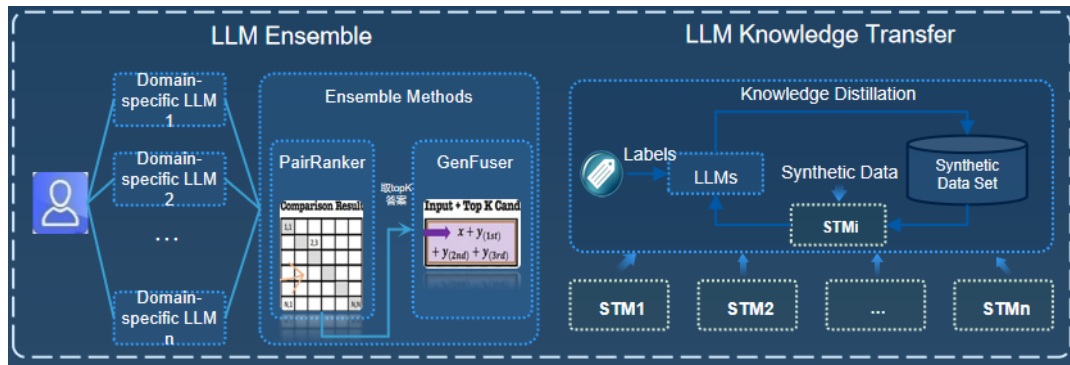

Figure 7-1 Augmented-Cognition LLMs

7.2 Scenario-Based AI Capability Lifecycle Support for Enterprises

TAC MaaS is capable of supporting AI capability lifecycle based on scenarios for continuous monitoring and optimization. According to ModelOps standards and best practices, it helps different roles in the enterprise to finely tune the LLMs/SLMs.
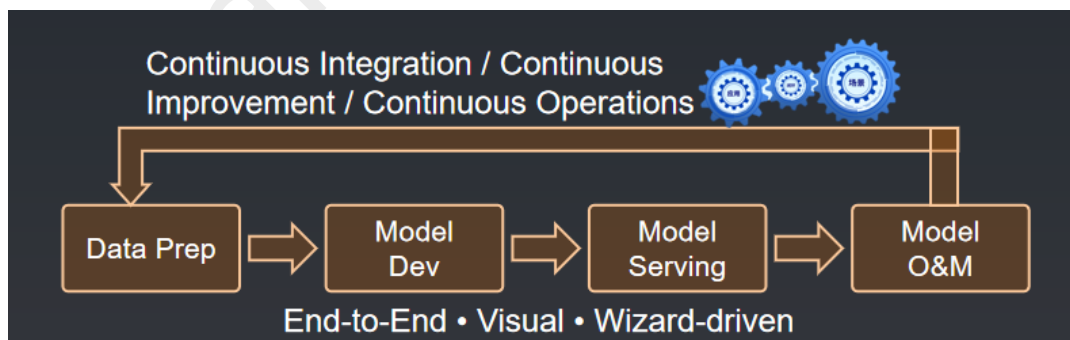

Figure 7-2 Scenario-Based AI Capability Lifecycle Support for Enterprises

---

7.3 LLM Guardian

TAC MaaS covers LLM guardian lifecycle from data processing, model training, risk evaluation, model deployment, to input/output compliance.

Figure 7-3 LLM Guardian

7.4 Diversified Deployment Solutions

TAC MaaS provides three types of composable and flexible solutions for private deployment demands.
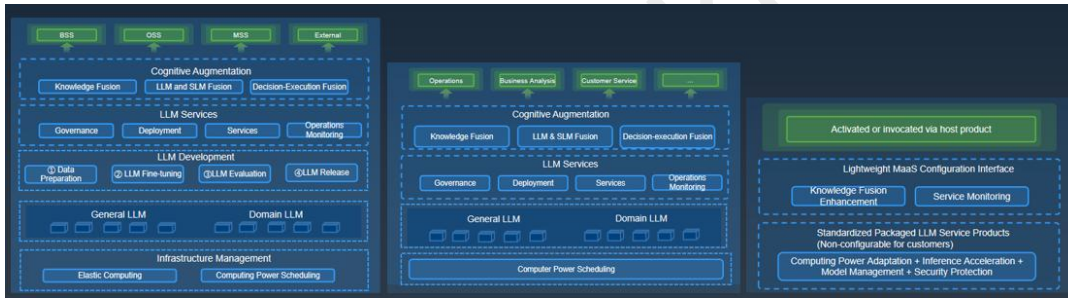

Figure 7-4 Diversified Deployment Solutions

8 Scenario Solutions

8.1 Solution for End-to-End LLM Foundation

8.1.1 Application Scenarios

This solution can be applied across telecom operators, finance, government, and transport with TAC MaaS serving as a foundational AI platform to unify the model hosting and resource scheduling.

8.1.2 Business Requirements

MaaS is now critical for industry-specific model development and service support. But there still exists a gap between the LLM and industry demands.

Compared with traditional development workflow, developing AI services is complex with long cycle but low efficiency. Simultaneously, AI service O&M is segmented in implementation, resources, and workflows, leading to difficult management.

With digital intelligence transformation advancing, data, algorithms, computing power, and models are now serving as the core company assets, requiring a unified AI infrastructure and collaboration mechanism.

8.1.3 Solution

This solution is aiming at providing customers with end-to-end integration capabilities in a unified AI platform from preparing data, developing/fine-tuning models, evaluating, managing, deploying, to open service. With accumulated data and algorithms, it builds up diversified development tool chain to cover LLM life cycle.
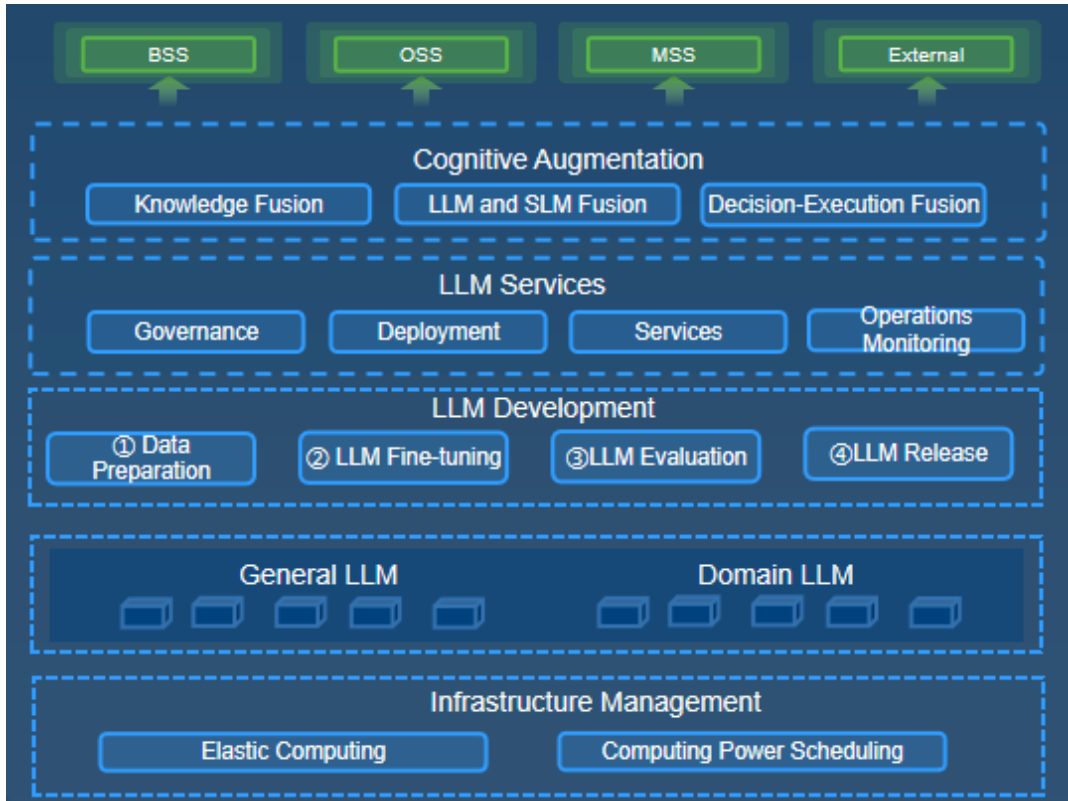
Figure 8-1 End-to-End LLM Foundation

8.2 Solution for LLM Middle Office

8.2.1 Application Scenarios

This solution can be applied across telecom operators to unify the model management, augment cognition capabilities, and schedule the resources in a centralized LLM application platform.

8.2.2 Business Requirements

With increasing demands for LLMs, parallel supplies from different vendors lead to segmented development, complex O&M, and low efficiency. There are several obstacles:

• Siloed development gathers different AI services with various standards and structures;

• Low efficiency in docking services to different AI model/capabilities across systems;

- Uncentralized hardware resources lead to low utilization;

- Current platform tightly couples training and reasoning with low performance in compatibility; besides, this kind of platform can only be deployed for large enterprises but not suitable for mid-/small-size companies.

8.2.3 Solution

Based on an open AI reasoning platform, this solution is building an intelligent AI foundation with AI toolchain for enterprises to support multi-source LLMs and schedule multi-structure resources. It can realize model sharing, quick deployment, and customized application development.
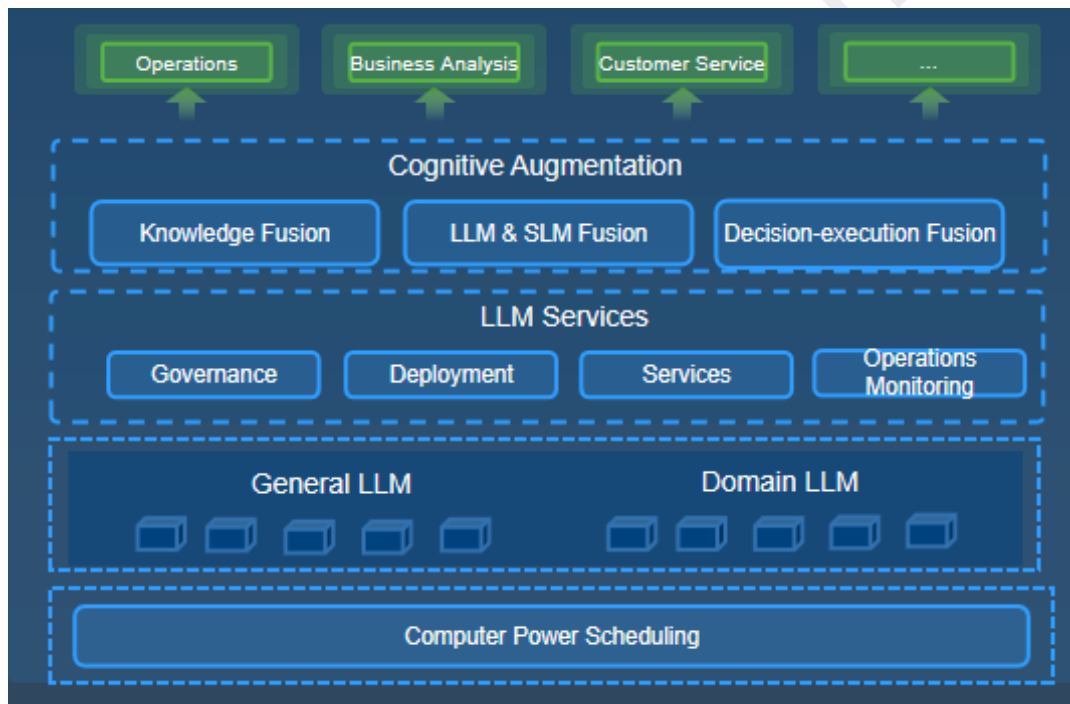

Figure 8-2 LLM Middle Platform

8.3 Lite Embedded MaaS Solution

8.3.1 Application Scenarios

This solution can be applied across telecom, finance, governments, and construction sites by augmenting legacy systems with LLMs and knowledge, accelerating agile application development and intelligence upgrades.

8.3.2 Business Requirements

When upgrading existing product/application or developing new applications, the medium-/small-size companies require lightweight LLMs. On the one hand, the AI platforms from most vendors are complex that are difficult to quickly deploy and integrate; on the other hand, the customers' hardware resources are limited to support such complex platform.

8.3.3 Solution

This solution gathers adaptive computing power, model management, accelerated reasoning, and safeguards to provide customers with standard LLM services and tools. Customers can integrate LLM capabilities quickly when upgrade products and develop applications.
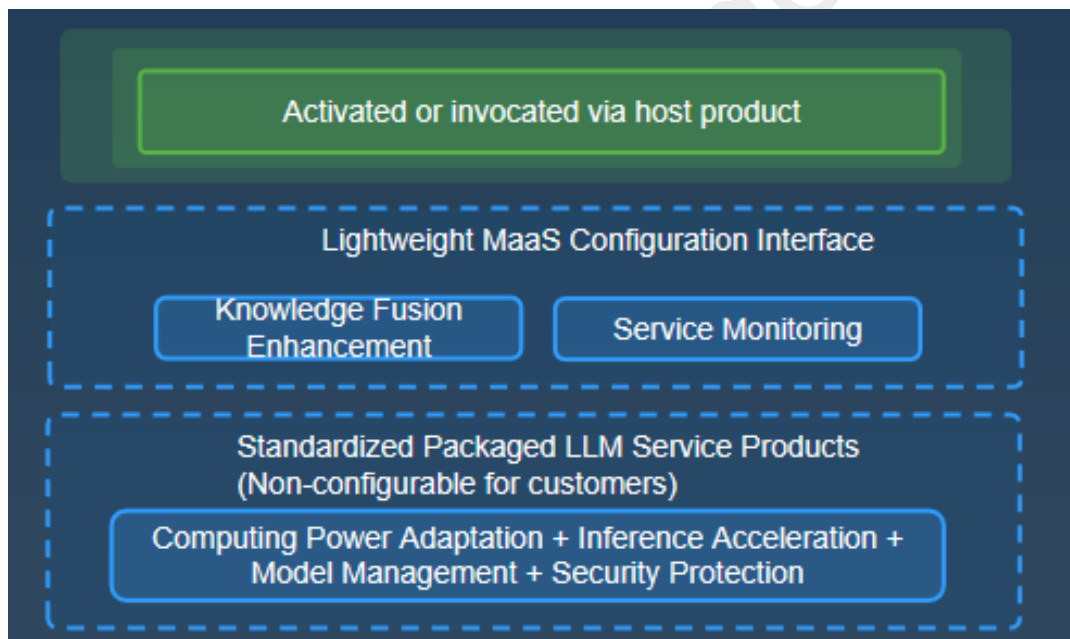

Figure 8-3 Lite Embedded MaaS Solution

8.4 One-Stop AI Development Platform

8.4.1 Application Scenarios

This solution can be applied across telecom, finance, governments, and construction sites by empowering unified AI infrastructure and realizing AI asset sharing.

8.4.2 Business Requirements

AI development contains various algorithms and frameworks, requiring comprehensive developer skills; simultaneously, there still exist bottlenecks such as complex developing/data environments and unconnected workflows.

Customers are seeking a unified platform and management architecture to gather resources and enhance O&M, as well as to transform data, algorithms, computing power, and models into company assets.

8.4.3 Solution

This solution guides users with a pipeline workflow across R&D from preparing data, developing models, to deploying. It is preset with rich capabilities to link systems and data chains in customer's system to reduce tool installation, data migration, and platform adaption.
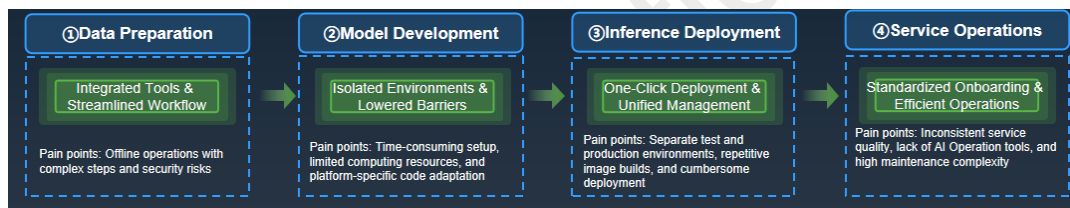

Figure 8-4 One-Stop AI Development Platform

9 Use Cases

TAC MaaS has served over 60 customers across industries with end-to-end solutions and services for digital intelligence upgrades, spanning customer groups to non-telecom domains, such as finance, energy, public security, media, transport, and governments.

9.1 An LLM Foundation for a Telco

9.1.1 Customer Requirements

The telco is aiming to empower its AI middle office with LLMs and building FMOps systems to enhance service efficiency.

9.1.2 Solutions and Effects

Based on TAC MaaS, its AI middle office upgrades to MaaS platform with several unique improvements:

- Unified multisource resource control: Hosting 64 P40s and 128 V100s, totaling 192 GPUs efficiently and dynamically with computing power utilization rate increased by more than 60%;

- Unified LLM foundation: Hosting Vicuna-33B, ChatGLM2-6B and other open source LLMs, and establish more than 10 fine-tuning models; supporting intelligent applications such as AI Q&A, business analysis, data collection, and O&M.

- LLM FMOps system: Developing end-to-end operations systems from corpus preprocessing, multiple supervised fine-tuning, augmented domain knowledge, LLM+SLM, to augmented decision execution.
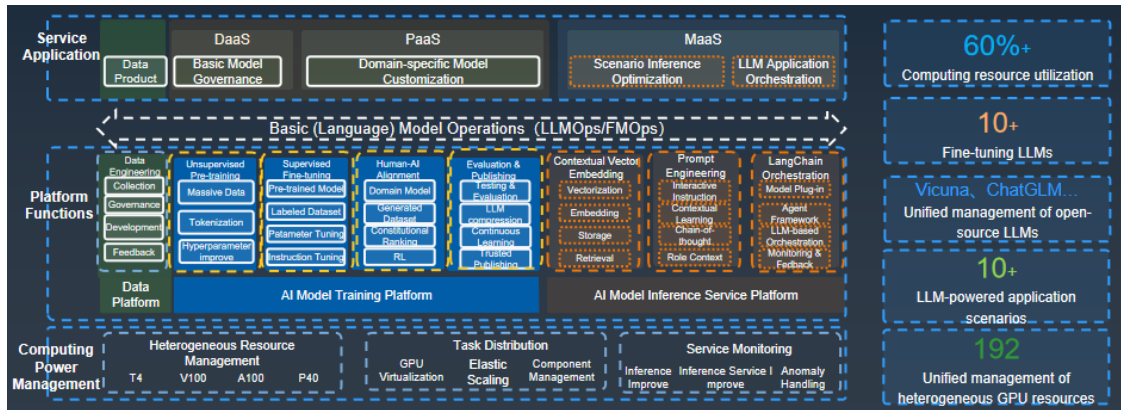
Figure 9-1 An LLM Foundation for a Telco

9.2 A Smart Data Platform for a Regional Telco

9.2.1 Customer Requirements

Siloed AI development causes segmented workflows/resources and poor data protection, so the telco is seeking an end-to-end AI middle office to unify data standards, service scheduling, and authentication management.

9.2.2 Solutions and Effects

Based on TAC MaaS, the telco has upgraded its middle office with data and technologies and enables multiple deployment for B/O/M domains, empowering all-round intelligence from perception, cognition, to decision-making.
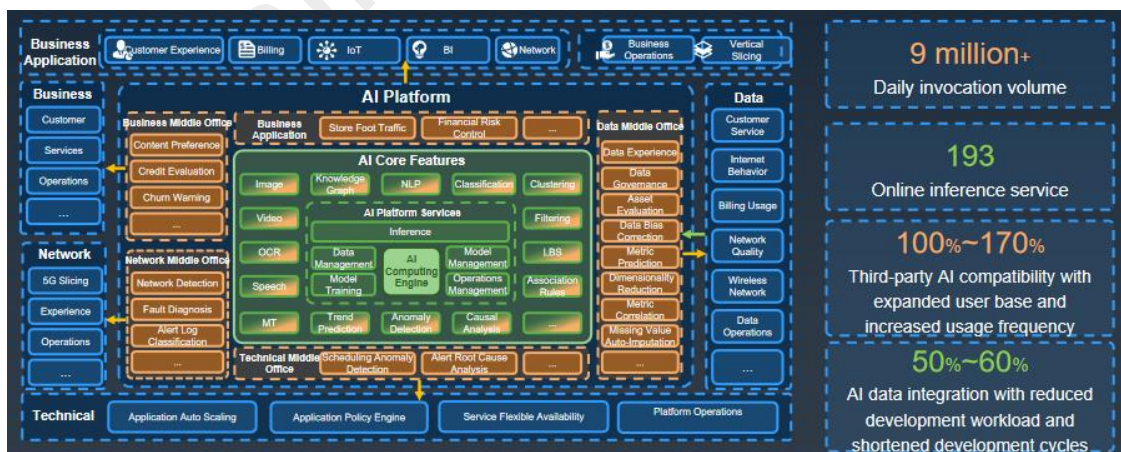

Figure 9-2 A Smart Data Platform for a Regional Telco

10 Certificates and Awards

10.1 Certificates

Part of certificates is shown in 错误!未找到引用源。:

Table 10-1 Part of Certificates

No.	Certificates
1	2022 AISWare AI² selected for MIIT's Key Innovation Tasks in AI
2	2022 AISWare AI² Edge Intelligence Computing All-in-One Machine certified by CCC and CNAS software testing
3	2022 the federated learning platform was certified as a member of the FATE open-source community, and Yuan Zhiyong was elected as the chairman of the FATE community multi-platform SIG
4	2024 28 compatibility certifications with domestic platforms including Hygon, Huawei, Inspur, etc.
5	2024 AISWare Usights registered with ICP (Record No. 110108442033101240017)

10.2 Awards

A selection of the awards the product has received is shown in 错误!未找到引用源。:

Table 10-2 Part of Awards

No.	Awards
1	2022 National Champion Grand Prize, China RPA+AI Developer Competition
2	2022 Wu Wenjun AI Science and Technology Award
3	2022 First Prize in CCF Science and Technology Progress Award

No.	Awards
4	2023 Second Prize in Beijing Science and Technology Progress Award
5	2023 Trusted AI: Outstanding Case in Generative AI Applications

10.3 Patents

Patents are presented in 错误!未找到引用源。:

Table 10-3 Part of Patents

No.	Patents	International/China
1	AI Resource Allocation Method Based on Signaling Interaction	International
2	Determination Method, Device, and Computer Equipment of User Activity Area Similarity	China
3	Method for Second-Level Data Anomaly Detection Using Dynamic Threshold Vector Operators	China
4	Method for Strategic Attack-Defense Decision Evaluation	China
5	End-to-end License Plate Vertex Localization and Recognition Based on Deep Learning	China
6	Batch Alert Root Cause Analysis Based on FP-Growth	China
7	Traffic Congestion Early Warning Using Complex Topology Analysis	China
8	Portable Device for User Identity Information Collection and Verification	China
9	Model Optimization Methods based on Metric Analysis Scenarios	China

No.	Patents	International/China
10	Narrowband Interference Suppression in UWB IoT Using High-Order Signal Processing	China
11	Training Method and System for Named Entity Recognition	China
12	Domain Knowledge-based LLMs Q&A Method and Device	China

10.4 Industry Standards

AsiaInfo is participating in several global and domestic standards, shown in 错误!未找到引用源。.

Table 10-4 Some Industry Standards

No.	Standard
1	IEEE P3398 Recommended Practice for GPT Empowered Software Engineering Life Cycle
2	Technical Requirements and Evaluation Methods for Intelligent Agents by CAICT
3	General Capability Requirements for AI Development Platforms by AI Industry Development Alliance
4	Reference architecture, functional requirements, and performance testing methods for AI models and services by CESIAIIA: General Capability for AI Development Platforms – Part 4: LLM Technical Requirements

11 Contact Us

AsiaInfo Technologies (China) Ltd.

Address: AsiaInfo Building, 10 East Xibeiwang Rd, Haidian District, Beijing, China 100193

Postcode: 100193

Fax: (+86) 010-82166699

Tel: (+86) 010-82166688

Email: 5G@asiainfo.com

Web: www.asiainfo.com

Thank you

Customer Value Innovator & Digital Transformation Promoter with Full-Stack Data Intelligence