XFUSION

Let computing serve you better

FusionOne AI
Enterprise Token Factory Solution

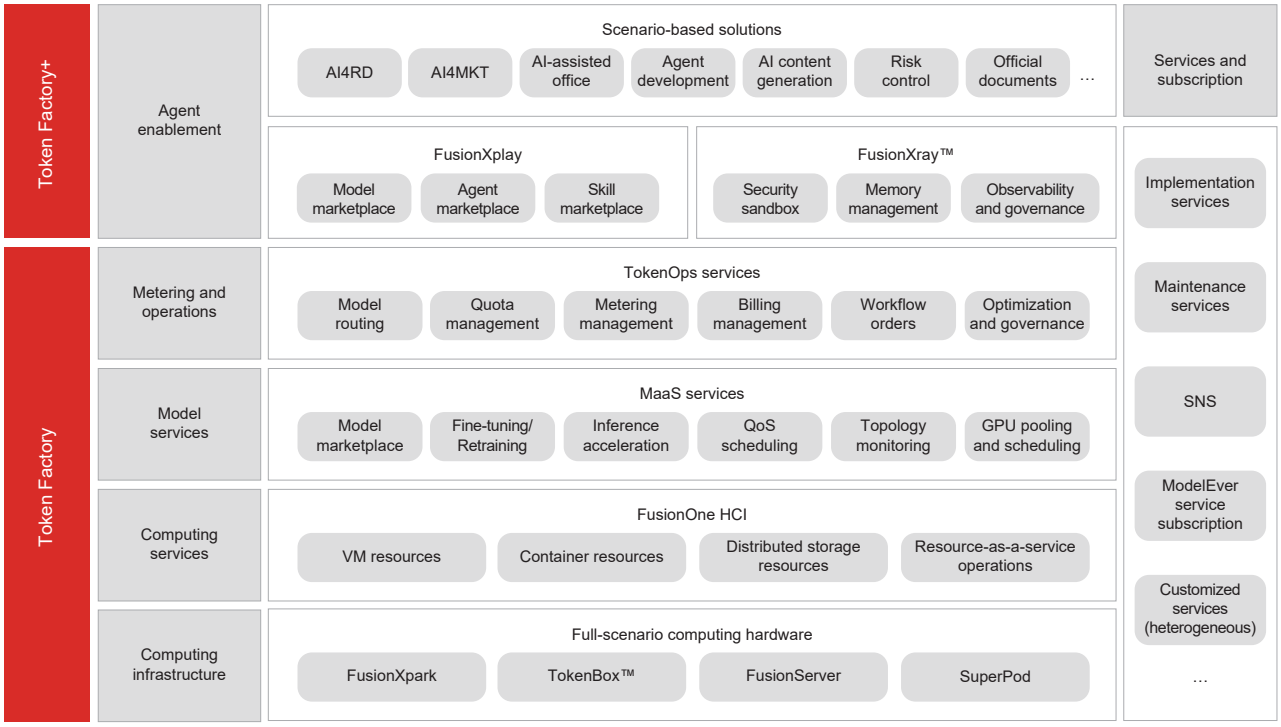
Dedicated Token Factory for Every Enterprise



XFUSION INTERNATIONAL PTE. LTD.

01 FusionOne AI
Enterprise Token Factory Solution

xFusion FusionOne AI is an enterprise-grade private AI token factory designed to maximize token efficiency across generation, utilization, and operations. At the computing layer, it provides a heterogeneous computing foundation spanning from edge to data center, enabling flexible AI deployment at scale. At the model service layer, FusionOne AI boosts inference performance through system-level inference acceleration across computing, storage, and networking, while enabling continuous model evolution. At the operations layer, a unified metering and operations platform delivers full visibility into token consumption and helps optimize usage efficiency. FusionOne AI also includes an agent enablement platform with a secure runtime environment and advanced memory capabilities, ensuring that agents run securely and accurately. Together with ecosystem partners, xFusion delivers integrated AI solutions tailored to industry scenarios, helping customers accelerate AI adoption and unlock greater business value.



xFusion's intelligent transformation practices

1 million+
Online meetings summarized

100,000+
Work orders dispatched and processed

100,000+
Financial documents processed

40,000+
Contract orders processed efficiently

80%
Self-service resolution rate

100%
AI-assisted coding adoption rate

6.2 billion+
Average daily coding tokens per user

56
Coding AI agents released

157
Skills covering requirements, design, coding, testing, and delivery

3190 lines/ person-month
Code output per engineer

44%+
End-to-end code generation rate

FusionOne AI — Enterprise Token Factory Solution

02 Driving AI Adoption with Leading Industry Partners

FusionOne AI

Driving AI adoption with industry partners

AI agent development platform

- Q&A
  Contract review
- Customer service training and assessment assistant
- Automatic email drafting
  Automatic PPT generation

Government AI agent

- Authoritative corpus-based knowledge bases
  Official document writing and proofreading
- Policy inquiry
  Meeting assistant
- Customer service agent assist
  Hotline early warning

Education AI agent

- Zero-code course development
  AI question generation and grading
- Pre-built practical training templates
  Programming courses
- Multi-discipline capability assessment and certification

Supply chain AI agent

- Production scheduling
  Planning optimization
- Inventory optimization
  Vehicle scheduling
- Product forecasting
  Shift scheduling optimization

Financial AI agent

- OCR order review
  Due diligence
- Financial and operating risk analysis
- Outbound calling
  Training companion

Healthcare AI agent

- Medical triage
  Imaging assistance
- Clinical decision support
  Medical record generation
- Quality control
  Screening

Proven across 1000+ customer deployments

No. 1 in China's AI appliance market, *source: IDC, China AI Appliance Market Share Report, 2025

National Medical Artificial Intelligence Pilot-Scale Facility

Healthcare AI innovation platform

- Inconsistent diagnostic capabilities across primary care institutions
- Complex TCM ingredients and lengthy R&D cycles

92%
Clinical accuracy with TCM four-diagnosis analysis

50%
Faster target prediction and compound screening than traditional methods

FIYTA

Precision manufacturing

Smart wearable AI platform

- Large number of hardware stock keeping units (SKUs) and lengthy R&D cycles
- Difficulties in integrating large volumes of multi-sensor data

80%
Greater R&D efficiency through hardware-assisted design

30%
Faster response from smart wearable assistant

Qilu Pharmaceutical

Biopharmaceuticals

Pharmaceutical DeepSeek application platform

- Long drug trial period
- Large amount of information in drug research literature, with low analysis accuracy

40%
Improved accuracy in drug target discovery

70%
Improved accuracy in scientific literature and document analysis

China Railway Chengdu Group

Transportation

Intelligent digital platform

- Complex segment and vehicle scheduling
- High customer service workload

15%
Improvement in punctuality through vehicle scheduling optimization

30%
Improvement in customer satisfaction through intelligent customer service

03 Six Core Capabilities for Faster Token Factory Adoption

FusionOne AI

Efficient token generation and utilization

Agent enablement

⑤ Deployable

Secure and governed agent execution

- Secure sandbox, memory management, and knowledge management
- Cross-agent memory sharing for less redundant inference, and semantic retrieval and memory orchestration that improves task execution accuracy by 30%

⑥ Deliverable

Accelerated AI adoption

- AI Lab with 152 PFLOPS of computing, supporting over 75 AI accelerator card models and rapid validation of computing and model resources
- FusionXplay AI marketplace with pre-built models, agents, and skills

Metering and operations

④ Governable

Consumption visibility

- Token consumption statistics by scenario, organization, and model, with support for cost allocation

Cost optimization

- Requests routed to the most cost-effective model based on scenario complexity

Budget control

- Built-in budget control, rate limiting, circuit breaking, and compliance auditing

Computing and model services

② Performant

Software-hardware synergy for accelerated inference

- 100+ tuning engineers and optimization across 500+ models
- 30% greater token inference performance

Model SmartQoS

- Intelligent request identification and QoS-based prioritization, ensuring stable concurrent access to models

Full-stack AI observability

- Unified view across the full stack from models to accelerator cards, with intelligent fault diagnosis

③ Adaptive

Day-0 model readiness

- Innovative decoupled architecture and automated adaptation pipeline
- Adaptation to latest models within days

Computing infrastructure

① Buildable

- Unified computing foundation from center to edge, delivering a standardized enterprise token generation pipeline
- Compatibility with over 75 accelerator card models and adaptation to new accelerator cards within days
- Unified management of third-party resources through a single access point, reducing O&M costs
- XPU Engine, enabling visualized management of heterogeneous computing and pooled scheduling at 1% granularity

04 Key Features
FusionOne AI

Key features

Resource Pooling	- XPU Engine enables fine-grained pooling and management of GPU and NPU resources. It supports multiple partitioning modes, including passthrough, vNPU, MPS, MIG, and vCUDA, with computing partitioning as fine as 1% and memory partitioning at the MB level. Multiple workload scheduling policies, including performance-aware and topology-aware scheduling, help optimize resource allocation and application performance. - Built-in visualization plug-ins provide chart-based monitoring of intelligent computing infrastructure. They cover GPU and NPU resource allocation, utilization, temperature, power consumption, and other key operating metrics. The platform also enhances alarming and monitoring for intelligent computing workloads on top of standard monitoring for traditional computing resources.
MaaS Platform	- Built on AI and cloud-native technologies, the one-stop platform manages the full lifecycle of AI models. It integrates data management, model development, training and fine-tuning, a model marketplace, model inference and deployment, model evaluation, model quantization, image management, and visualized O&M. It also incorporates a self-developed large model inference acceleration engine that improves inference performance, generates more tokens, reduces inference costs, and accelerates AI application deployment. - The platform connects to heterogeneous computing resource pools and is compatible with the self-developed high-speed inference engine Wings-Infer and leading open-source inference frameworks such as vLLM and SGLang. By deeply integrating acceleration technologies including adaptive tuning templates, operator acceleration, KV cache optimization, intelligent sparsity, speculative decoding, PD-separated scheduling, and large-scale expert parallelism (EP), it delivers low-latency, high-throughput model inference and ensures millisecond-level responses.
Metering and Operations	- The multi-tenant platform for fine-grained computing operations and governance supports quota control, tiered billing, and full-chain consumption traceability across the entire token usage lifecycle. All-perspective observability and traceability dashboards aggregate key token metrics in real time, including account balance, call frequency, token consumption, and average requests per minute (RPM) or tokens per minute (TPM), enabling comprehensive monitoring of services. - The platform supports metering and billing for a wide range of hardware resources, including CPUs, dual in-line memory modules (DIMMs), and accelerators, across full AI business scenarios, covering inference services, training services, online development, model quantization, evaluation, and so on.
Large Model Application Generation Suite	- The optimized retrieval-augmented generation (RAG) workflow powered by large language models (LLMs) enables accurate question answering and citation generation across data in various complex formats. Users can quickly build AI applications based on large models by low-code development and integrate these applications into their own business systems through APIs. - Application frameworks such as LangChain-Chatchat, FastGPT, and Dify are supported. - Vector databases such as Faiss and OpenSearch are supported.
O&M	- Hybrid deployment and unified GUI-based management of x86 and ARM servers. - AI-based memory fault self-healing and disk fault prediction, ensuring data security and minimizing server downtime. - Physical server monitoring with faulty disk indicators and diverse computing resource visualization. - Wizard-based installation, GUI-based configuration, and one-click system initialization. - Unified resource management, performance monitoring, alarm management, operation log management, permission management, hardware management, health checks, and log collection. - One-click scaling, log collection, and health inspection.

For more information about xFusion,

contact the local representative office or visit our official website at

https://www.xfusion.com/en.



Copyright ©xFusion International Pte. Ltd. 2026. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of xFusion International Pte. Ltd.

Trademarks and Permissions

xFUSION and other xFusion trademarks are trademarks of xFusion International Pte. Ltd. All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

In this document, "xFusion" is used to refer to "xFusion International Pte. Ltd." for concise description and easy understanding, which does not mean that "xFusion" may have any other meaning. Any "xFusion" mentioned or described hereof may not be understood as any meaning other than "xFusion International Pte. Ltd.", and xFusion International Pte. Ltd. shall not bear any liability resulting from the use of "xFusion".

The purchased products, services and features are stipulated by the contract made between xFusion and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.