Fujitsu Kozuchi Enterprise AI Factory Introductory Materials

Fujitsu Limited

Market Trends Surrounding Generative AI

© 2026 Fujitsu Limited

Fujitsu's Approach

uvance

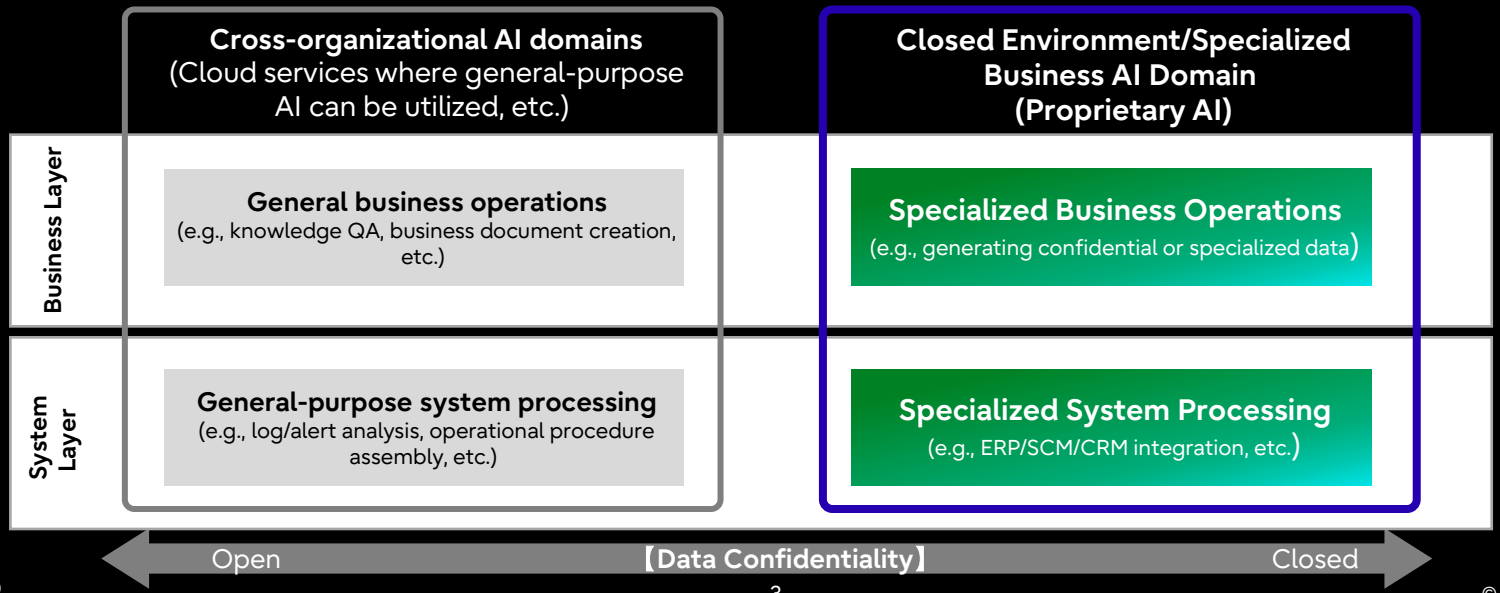
- Configuring a generative AI platform as a private environment capable of securely handling "highly confidential data" presents a new challenge

Risk of highly confidential data being taken outside the company

Ensuring an environment where operations can be proactively controlled

Optimization of AI models/agents tailored to specific business operations

→ Fujitsu's Generative AI Platform as a Private Environment
Generative AI Platform



© 2026 Fujitsu Limited

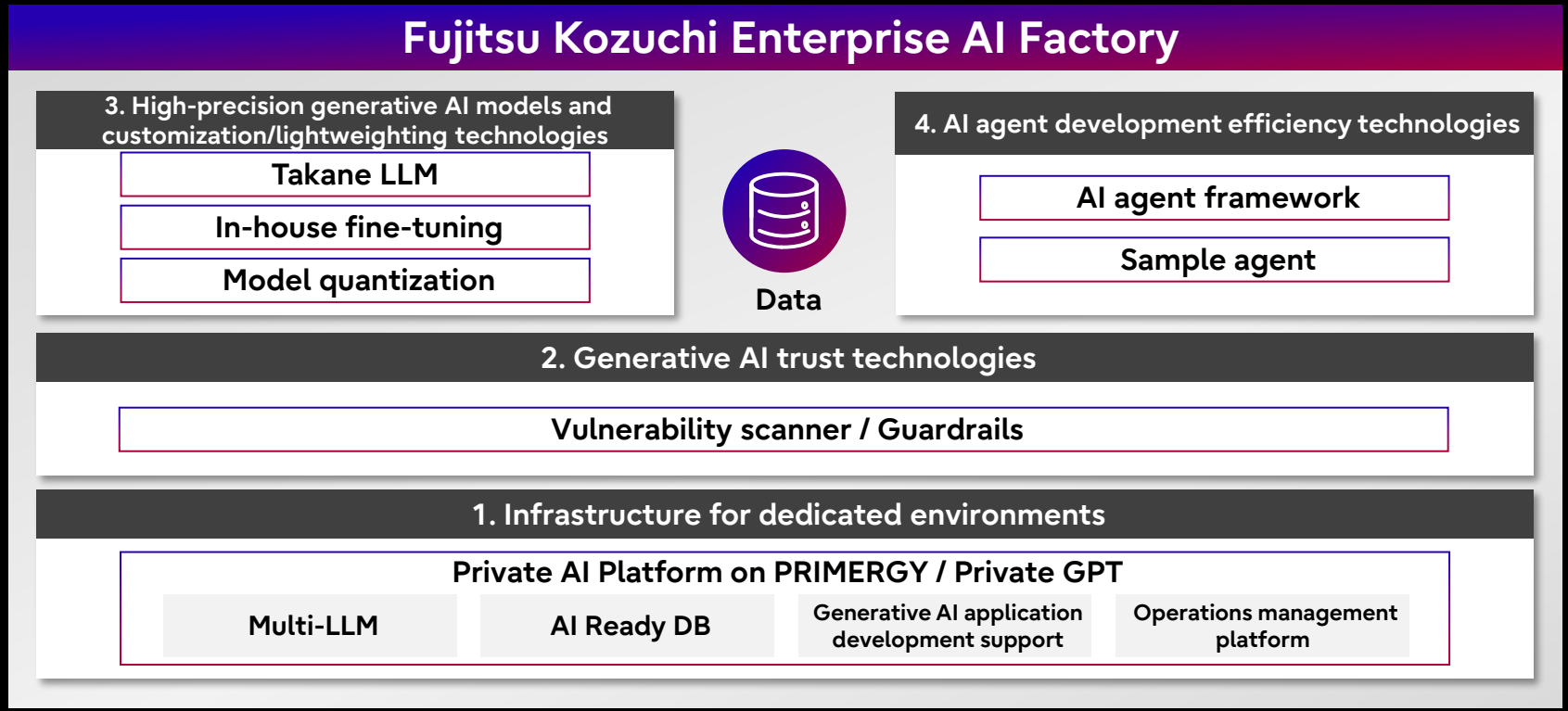
Fujitsu Kozuchi
Enterprise AI Factory Overview

© 2026 Fujitsu Limited

Fujitsu Kozuchi Enterprise AI Factory

uvance

- A proprietary AI platform that can be utilized on-premises as a means to achieve highly sovereign AI



© 2026 Fujitsu Limited

Fujitsu Kozuchi Enterprise AI Factory

uvance

- **A proprietary AI platform deployable on-premises as a means to achieve highly sovereign AI.**

1. Infrastructure for dedicated environments

- Achieving on-premise AI utilization that keeps data internal.
- Providing end-to-end support from introduction to advanced operation.
- Adopting PRIMERGY-based high-performance GPU infrastructure. *

2. Generative AI trust technologies

- LLM vulnerability scanner supporting over 9,000 types of vulnerabilities.
- Guardrails to prevent malicious input and unsuitable output.
- Automated operation ensuring safety even for non-experts.

3. High-precision generative AI models and customization/lightweighting technologies

- Adopting the high-precision Japanese LLM "Takane" as the core.
- Developing and refining domain-specific models with in-house fine-tuning.
- Achieving low-cost and highly efficient AI operation through LLM quantization technology.

4. AI agent development efficiency technologies

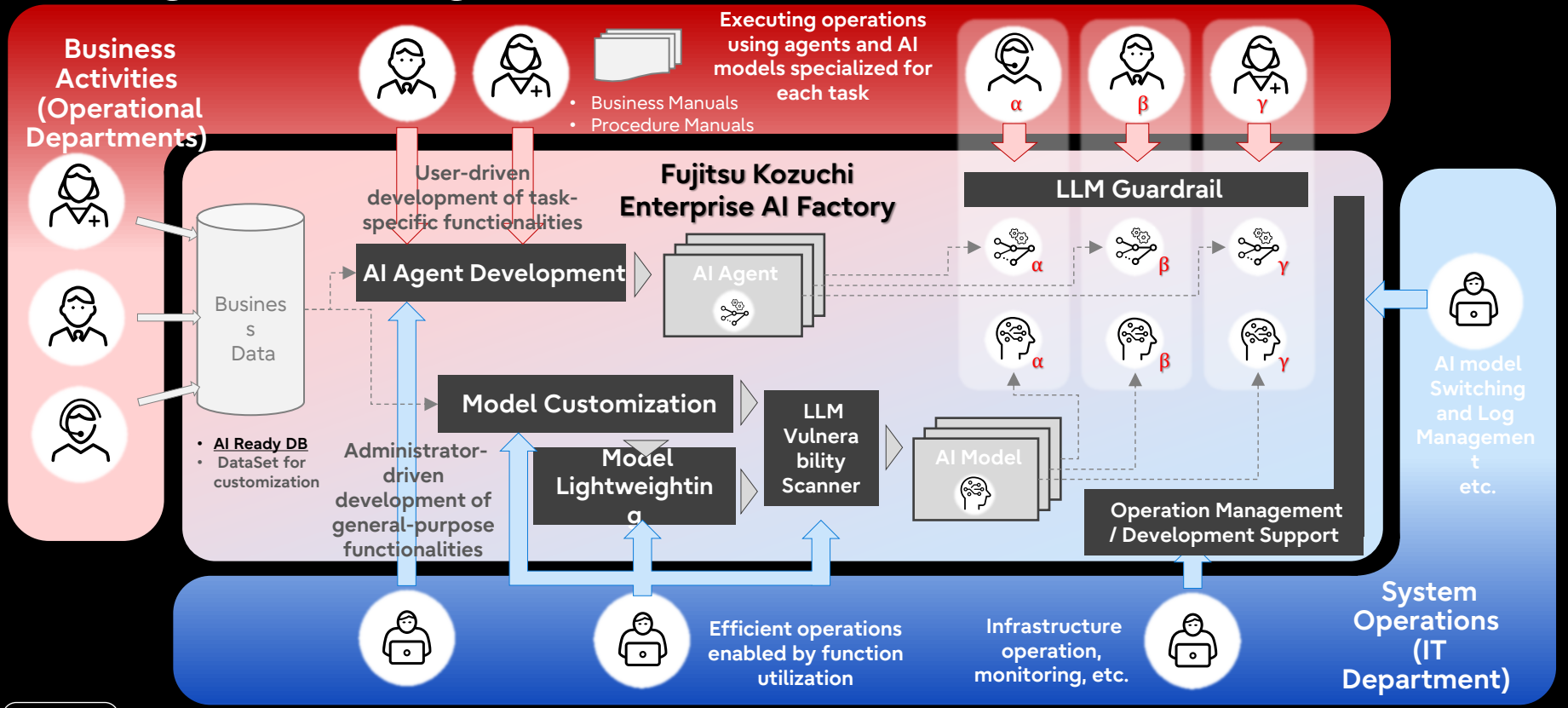
- Quickly developing enterprise AI with low-code/no-code platforms. Rapidly build business AI
- Flexible integration with existing systems through MCP compatibility.
- Advanced utilization through collaborative actions of multiple AI agents.

*Japan: Private AI Platform on PRIMERGY, Europe: Private GPT

© 2026 Fujitsu Limited

In-house production of specialized models driving business growth

uvance



© 2026 Fujitsu Limited

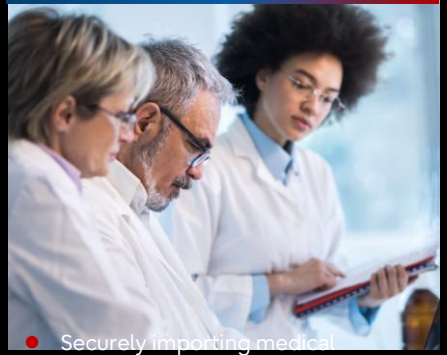
Key Target Areas

- **Our primary targets are set to four domains: Manufacturing, Healthcare, Finance & Insurance, and Social Infrastructure & Public.**
- **A proprietary AI platform that enables the development, operation, and continuous advancement of AI tailored to industry requirements.**

Manufacturing

- Cross-handling drawings, manuals, and equipment data generated in each stage of design, production, and maintenance, and iteratively deploying them to on-site operations.
- Accumulating and reusing knowledge related to design intent understanding, quality analysis, robotics, and automated equipment across all processes.

Healthcare



- Securely importing medical records, examination reports, and internal documents, and continuously utilizing them throughout the entire business process.
- Streamlining and enhancing healthcare operations through the standardization and reuse of document generation, search, and summarization, based on on-premise and closed environments.

Finance and Insurance

- Centralized management of business data such as policies, regulations, review materials, and inquiry histories, embedding them into business processes for continuous utilization.
- AI utilization predicated on stable operation and scalability, achieved through the standardization of inquiry responses and document verification, and the reflection and updates of business rules.

Social Infrastructure & Public Services

- Systematically managing administrative documents, operational manuals, and facility/operations information for iterative use across organizations.
- Sustainable operation based on closed and proprietary environments in the fields of disaster prevention, transportation, and energy, as well as national security and critical infrastructure domains.

© 2026 Fujitsu Limited

Fujitsu Kozuchi Enterprise AI Factory Details

© 2026 Fujitsu Limited

Function Stack (Configuration Overview)

uvance

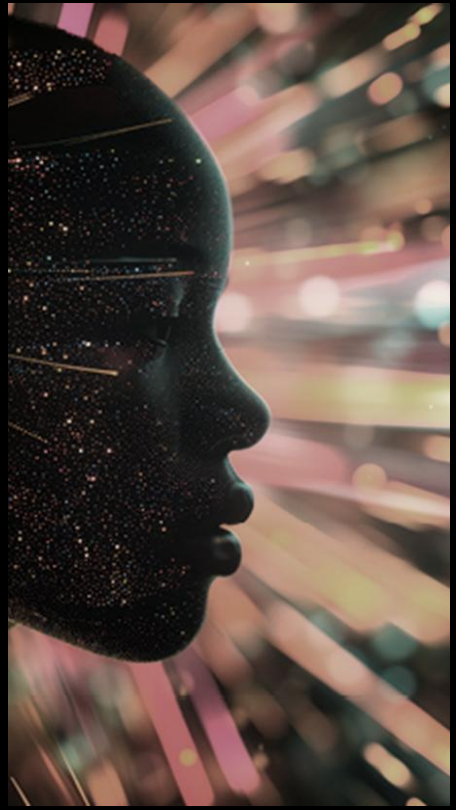
- **Providing all necessary functionalities for each layer (infrastructure, generative AI model, and application) in a single package.**

④ AI agent development efficiency technologies	Generative AI Applications Development Support	Agent Platform	Dify etc.	Code Execution	Jupyterlab	Real-time audio conversion	Pyannote.audio
				Log Management	Langfuse	Speaker Identification	Pyannote.audio
				Model Switching	Streamlit	Text-based Charts RAG	docling
③ High-precision generative AI models and customization/lightweighting technologies	Multi-LLM	Model (LLM/VLM /MLLM)	Takane	gpt-oss Gemma3/4 Nemotron	Embedding	Ruri	OpenAI Compatible	Lite LLM
					Rerank	Ruri Reranker	LLM Execution Platform	vLLM
	Model Customization & Lightweighting		Customizatio n	In-House Fine-Tuning				
			Lightweightin g	Model Quantization (Quantization PG / Operational Support PG)				
	AI Ready DB		pgvector					
② Generative AI trust technologies	Generative AI Trust		LLM Vulnerability Scanner & Guardrails					
① Infrastructure for dedicated environments	Operationnal Management	Metrics Monitoring	Prometheus	Metrics Collection	Node and GPU Exporter	Dashboard	Grafana	
	Infrastructure Platform	Container Infrastructure	Podman					
		OS	Red Hat Enterprise Linux (RHEL)					
		GPU Server	PRIMERGY					

© 2026 Fujitsu Limited

Private AI Platform on PRIMERGY[PAPP]

uvance



• Pre-built Conversational Generative AI Platform

A pre-built generative AI platform, finely balanced based on our AI engineers' practical knowledge, is delivered as a ready-to-use model.

• Installation/Maintenance Services

Reliable and smooth installation services by experienced engineers, along with hardware/OS maintenance services supported by our help desk, are provided.

• Related Services

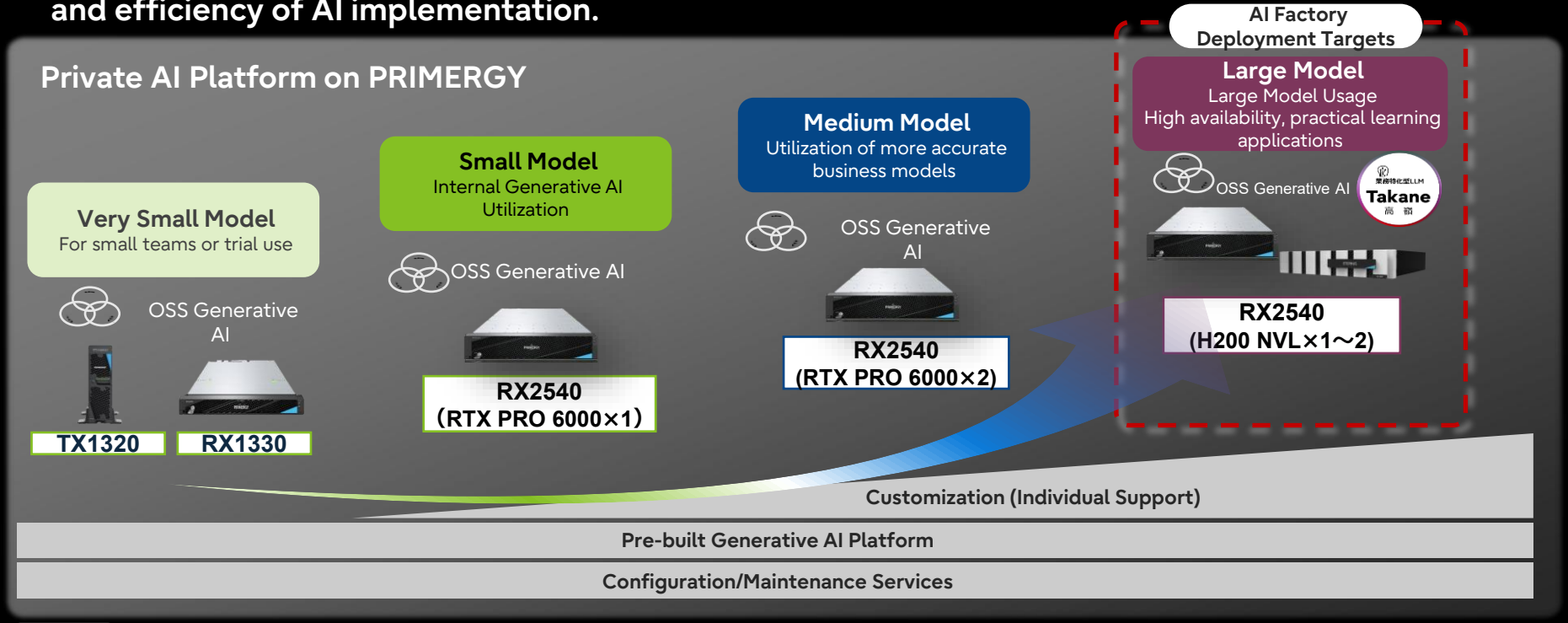
A variety of related services are offered to address diverse customer challenges, from pre-implementation assessment to post-implementation maintenance and operational support.

© 2026 Fujitsu Limited

Private AI Platform on PRIMERGY [PAPP]

uvance

- A diverse range of models, from Very Small to Large, is offered to suit various scales.
- Flexible configurations tailored to customer needs are possible, significantly enhancing the speed and efficiency of AI implementation.



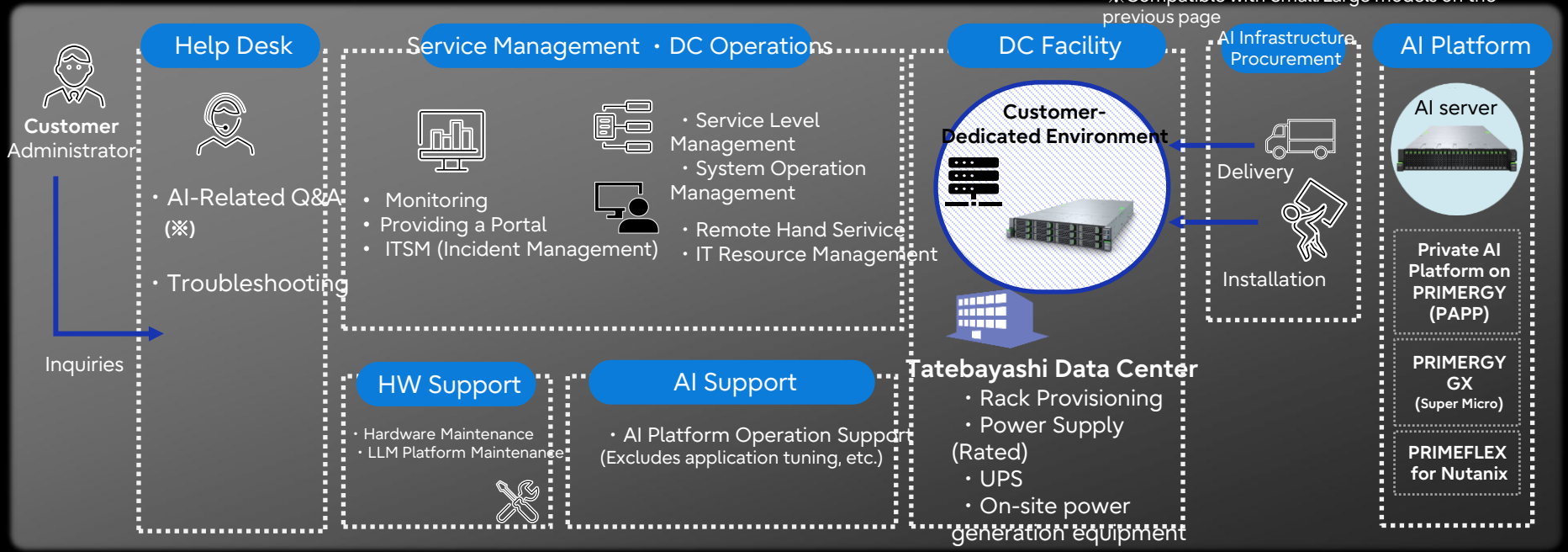
© 2026 Fujitsu Limited

On-Premises Generative AI Solution [Managed]

uvance

- Providing the necessary hardware (HW) and LLM environment for generative AI utilization, set up and delivered from Fujitsu's data centers.
- Configured as a dedicated environment, enabling use in a highly confidential setting without sharing with other companies.
- Offering full support with IT resource monitoring, covering troubleshooting upon malfunction and AI-related QA support.

※Compatible with Small/Large models on the previous page

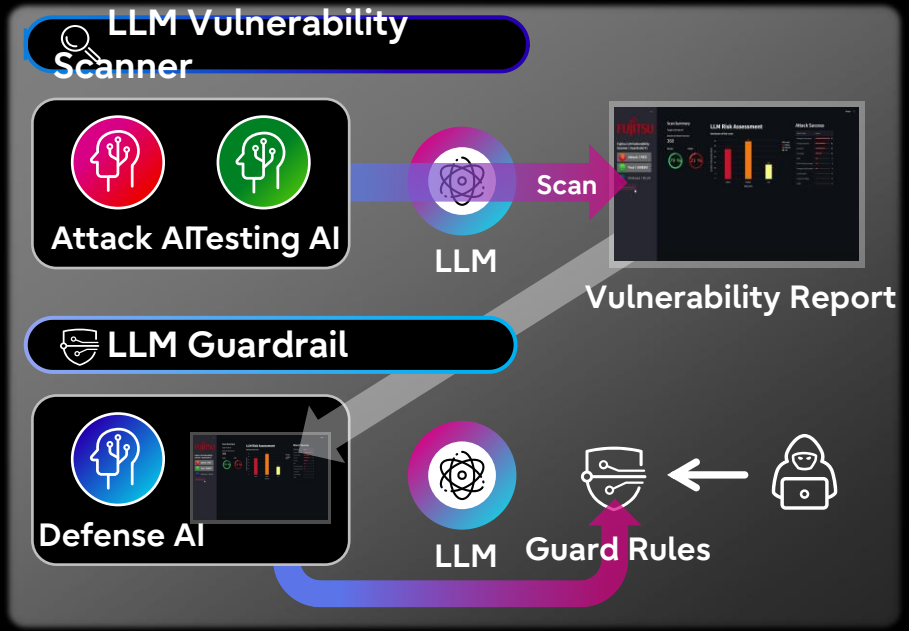


© 2026 Fujitsu Limited

LLM Vulnerability Scanner & Guardrails

uvance

- LLM vulnerabilities are identified through comprehensive evaluation, and optimal defense technologies are automatically applied.



Top-tier vulnerability coverage in the industry

- Supports over **9,000 types of vulnerabilities**, including Fujitsu's proprietary methods.
- Accurately analyzes **conversational attacks** that were previously difficult to detect.

Requires no expert knowledge

- Explanation features enable **non-experts** to easily understand the impact.
- Enables enhanced security measures with **reduced learning costs.**

Automatically generates defense rules

- **Automatically generates** optimal defense rules based on reports.
- Enables out-of-band defense **without altering the core LLM**

© 2026 Fujitsu Limited

LLM for Enterprise – "Takane"

cohere

- LLM optimized for Enterprise operations
- World-Class RAG Technology enabling high-precision use of in-house data
- Industry-leading Embed and Rerank models
- Highly customizable for key business tasks

Latest Large Language Model
「Command R+」

FUJITSU

- Japanese-specific additional training/Fine-tuning technology
- Composite AI technology
- Generative AI trust technology
- Provided in a private environment
- Knowledge accumulated through the building of mission-critical systems

© 2026 Fujitsu Limited

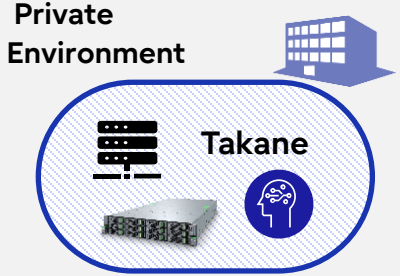
Why Takane?

uvance

✓ Enterprise specific high-secure LLM

Secure Data Utilization

Confidential on-premises data that cannot be moved to the cloud can still be safely leveraged in **a closed environment.**



AI That Knows Your Business

Customize LLMs to reflect your unique knowledge and deliver accurate, context-aware answers that strengthen competitive advantage.

Reliable Implementation Support

End-to-end assistance—from strategy to operations—ensuring smooth adoption of generative AI with minimal customer burden.

© 2026 Fujitsu Limited

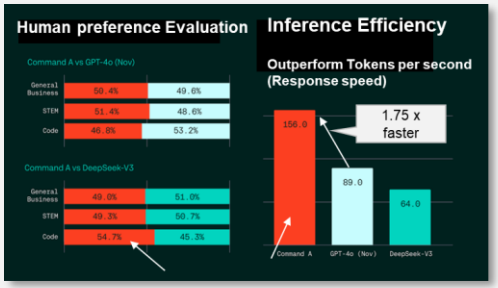
uvance

Why Takane?

✓ **Delivers high Japanese performance with minimal infrastructure resources**

High-Speed Inference

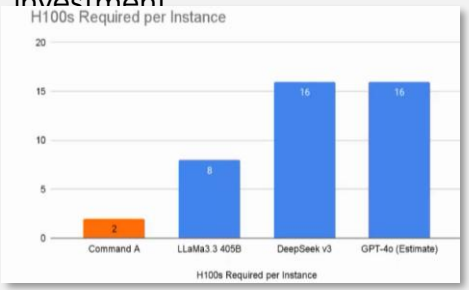
Delivers performance equivalent to general-purpose models while **achieving 1.75 times the output speed of GPT-4o**



Source: https://cohere.com/blog/command-a

GPU-Efficient

When using NVIDIA H100, **operates with 1/8th the resources** compared to GPT-4o. Delivers maximum performance with minimal infrastructure investment



Source: https://cohere.com/blog/command-a

Rich Lineup

Offering a diverse range of products from 1B to over 100B parameters

※2025.9 Company research

parameter size	Cohere/Takane	NTT tsuzumi	NEC cotomi	IBM granite	Ricoh	H2O
1-30B	Command R7B Takane 2.0-7B	tsuzumi 0.6B tsuzumi 7B tsuzumi 13B ※1	cotomi Fast v2 ※2	Granite 3.3 2B ※3 Granite 3.3 8B ※3		Danube3 0.5B Mississippi-800M_3 0.8B Danube2 1.8B Danube3 4B
30-100B	Command R Takane 2.0-32B		cotomi Pro v2 ※2		70B Llama 3.3 Ricoh-70B	
100B-	Command R+ Takane 1.1 Command A Command A MM				gpt-oss-120B	

Differentiation

※1 As of November 2024, it is being evaluated as the medium-sized version of Tsuzumi.
※2 The parameter size has not been disclosed.
※3 In addition to the language model, Granite also provides vision and code models.

© 2026 Fujitsu Limited

Takane Lineup

uvance

- **Select the appropriate model size based on your requirements**

Scale	LLM Products	Features	Use Cases
Small	**Takane 7B** Command R7B	• Small-scale Model • Quick Response • For small to medium-sized tasks	• Enterprise Chatbot • Content Generation • Automatic Translation
Middle	**Takane 32B** Command R	• Medium-scale model • Contextual Understanding • Complex Task Processing	• Research Institutions • Business Analysis • Advanced Customer Support
Large	**Takane** Command A	• Large-Scale Models • Advanced Language Processing • Specialized Domain Support	• Large-Scale Media • Specialized Fields • AI Assistant Development
Large Multi Modal	**Takane 112B Vision** Command A Vision	• Multi-modal Model • Reading Charts and Images	• Text Analysis Including Charts and Graphs • Reading Drawings in Design and Manufacturing Fields
Extension Model	**Embed 4**	• Maps text to a vector space, enabling similarity and semantic analysis	
	Rerank 3.5	• Reranks search results to quickly deliver highly relevant information	

© 2026 Fujitsu Limited

Diverse compatible models

uvance

OpenAI
gpt-oss 20b & 120b

OpenAI's high-performance Open-source LLM

A large language model that combines long input processing and complex contextual understanding, enabled by its massive parameters, with the flexibility for commercial use.

- gpt-oss-20b has 21 billion parameters, excelling in long-text processing and complex contextual understanding.
- gpt-oss-120b features 120 billion parameters, offering an accuracy comparable to GPT-4 while achieving both efficiency and precision.

Google
gemma3 12b & 27b
gemma4 26b & 31b

A multimodal LLM developed by Google

A lightweight, high-performance open-source AI model that can process images in addition to text.

- Generate text using not only text input but also image input.
- Gemma 3 and 4 can process images, while Gemma 4 can process videos as frame-by-frame images.
- Supports over 140 languages, including Japanese.

NVIDIA
Nemotron3 nano

A new LLM model developed by NVIDIA

A high-precision LLM that supports complex decision-making through long-context understanding and advanced reasoning.

- Supports long texts of up to 1 million tokens while delivering high-speed inference.
- By combining the Mamba architecture, which maintains efficiency even with long texts, with the MoE architecture, which reduces computational requirements, the model achieves significantly higher efficiency than previous models, further improved by technologies such as FP8-based quantization, GPU optimization, and NVFP4)

© 2026 Fujitsu Limited

Fine-Tuning & Model Quantization

uvance

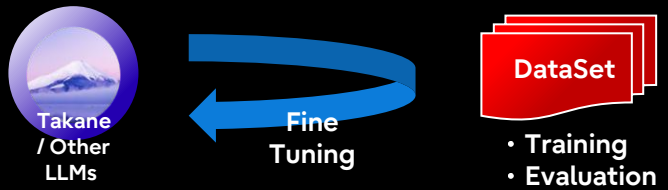
Fine-Tuning

- Low response accuracy
- Requires resources for business-specific customization
- Lack of personnel with specialized expertise

▶ **Build domain-specific models through our fine-tuning services**

Streamline fine-tuning execution and operations

If a training dataset is available Fine-tuning is possible



Enables the creation and improvement cycle of domain/business-specific models

Model Quantization

Generative AI consumes vast amounts of electricity, so it needs to be made lightweight according to its intended use

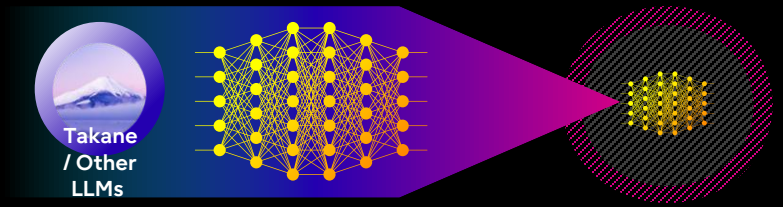
▶ **Achieved the world's highest accuracy with 1-bit quantization**

Quantization: A technique that compresses the weights assigned to connections between neurons
(Reduces model size by representing weights with fewer bits)

Achieving lightweight and low power consumption while maintaining accuracy

Before quantization (32-bit or 16-bit)

After quantization (1 bit)



Maintains high precision using Fujitsu's proprietary quantization error propagation method

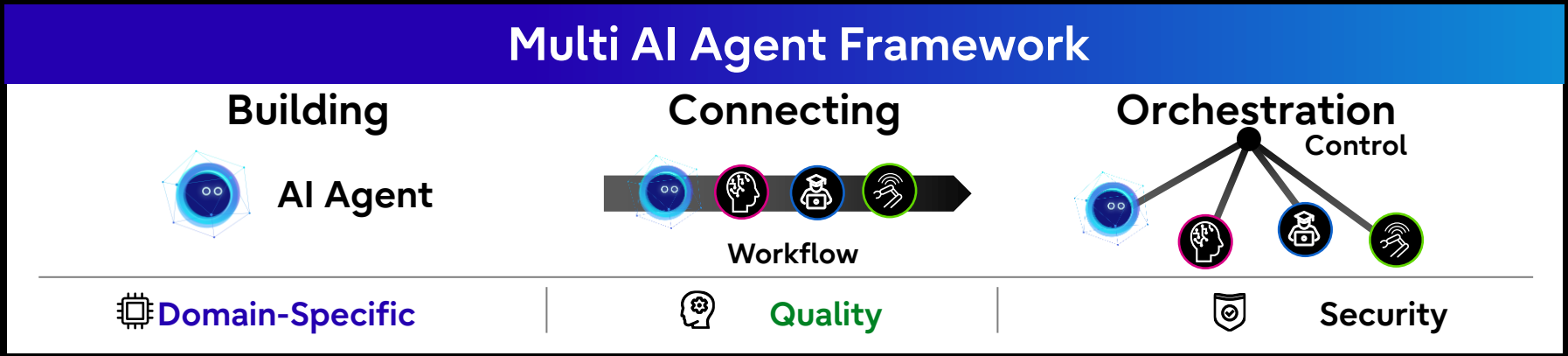
© 2026 Fujitsu Limited

Multi-AI Agent Framework

uvance

- Domain-Specific, High-Quality, Secure Workflow Automation for AI Agents

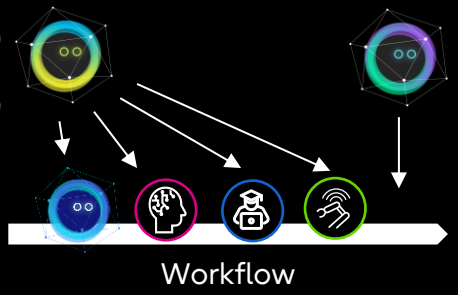
Multi AI Agent Framework



Domain-Specific

- Agentic Memory
- Integration

Acquires knowledge from humans, robots, and other agents.
Execute workflows while growing within specific domains.



Growth (Quality)

- Monitoring
- Constraint-Based Routing

Properly monitors complex interactions throughout the entire workflow.
Supporting workflow execution while considering diverse constraints.

© 2026 Fujitsu Limited

Use case

© 2026 Fujitsu Limited

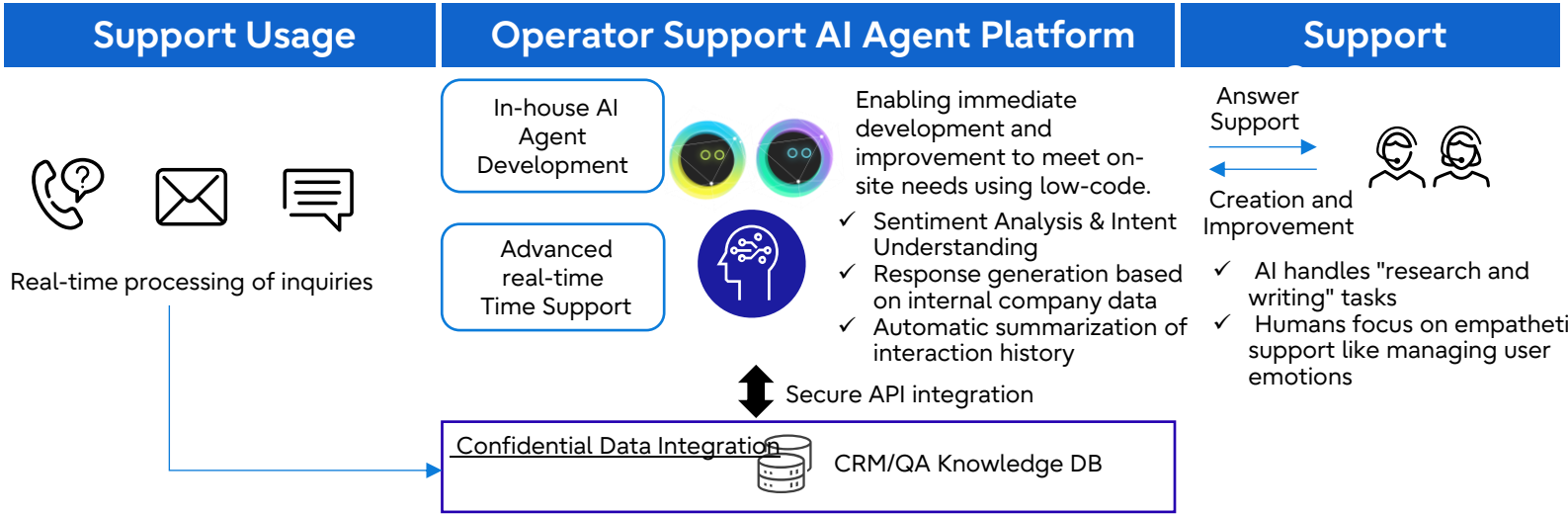
Advanced Customer Service

uvance

Background / Purpose

- Building a next-generation contact center platform for secure, low-code, field-initiated AI agent development

Utilization Usage Scenario



Expectations Effect

- Improving QCD (Quality, Cost, Delivery): Enhancing operator response quality and reducing handling time through the automation of search and recording tasks
- Business Transformation: Shifting support operator responsibilities from considering and creating responses to developing/improving agents and providing empathetic support to inquirers.

© 2026 Fujitsu Limited

Management Decision Support

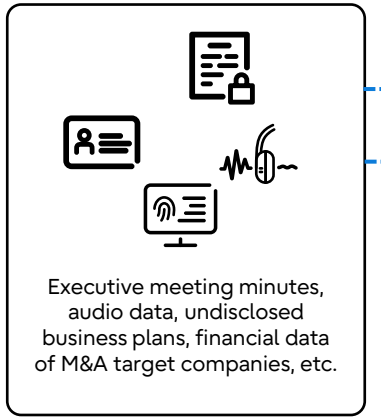
uvance

Background/ Purpose

- Extracting information from internal confidential documents and data sources to support executive decision-making

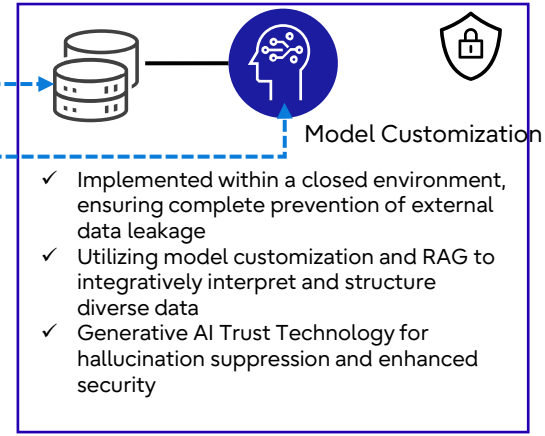
Application Usage

Diverse confidential information → A secure AI platform for learning and analysis → Advanced management decision-making



Executive meeting minutes, audio data, undisclosed business plans, financial data of M&A target companies, etc.

RAG



- ✓ Implemented within a closed environment, ensuring complete prevention of external data leakage
- ✓ Utilizing model customization and RAG to integratively interpret and structure diverse data
- ✓ Generative AI Trust Technology for hallucination suppression and enhanced security

Accelerate decision-making and improve accuracy

Market position forecasting, M&A evaluation, new business ideas

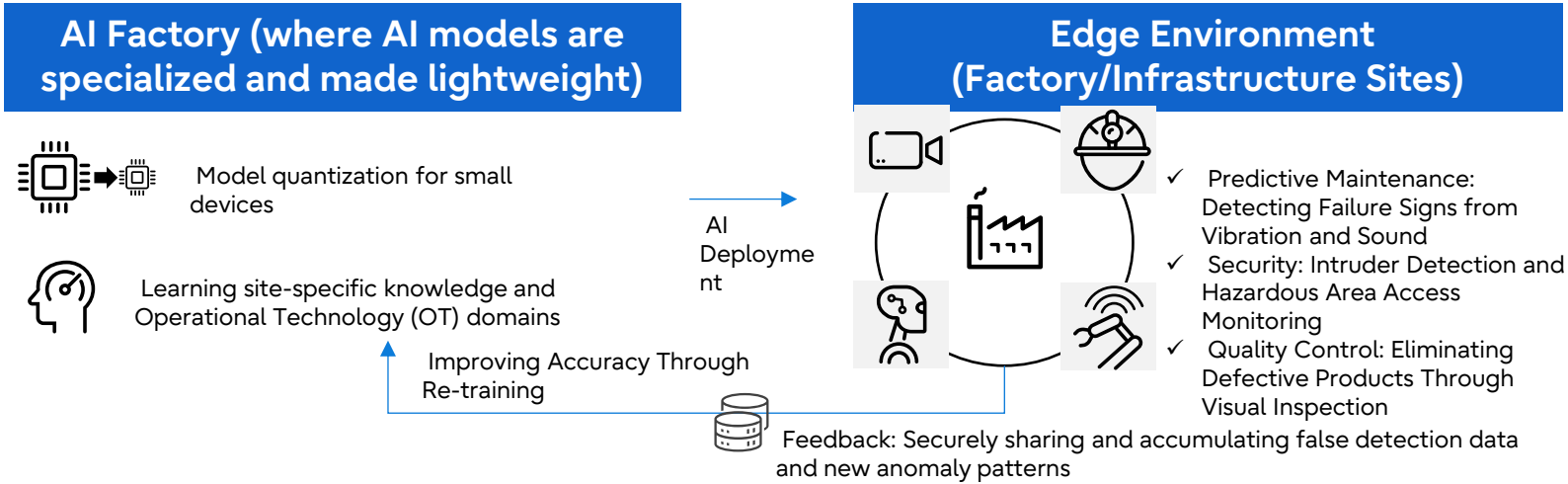
Expectations Effect

- Leveraging, analyzing, and inferring from data that can only be handled in a private environment, AI highly supports executive decision-making

© 2026 Fujitsu Limited

Edge AI evolving driven by manufacturing and infrastructure sites

uvance

Background/ Purpose	• Deploying lightweight models via quantization technology and in-house fine-tuning to cultivate AI specifically tailored for on-site operations
Application Image	
Expectations Effect	• Enhancing equipment utilization through predictive maintenance, automating visual inspections for consistent quality, and ensuring continuous monitoring of hazardous areas

© 2026 Fujitsu Limited

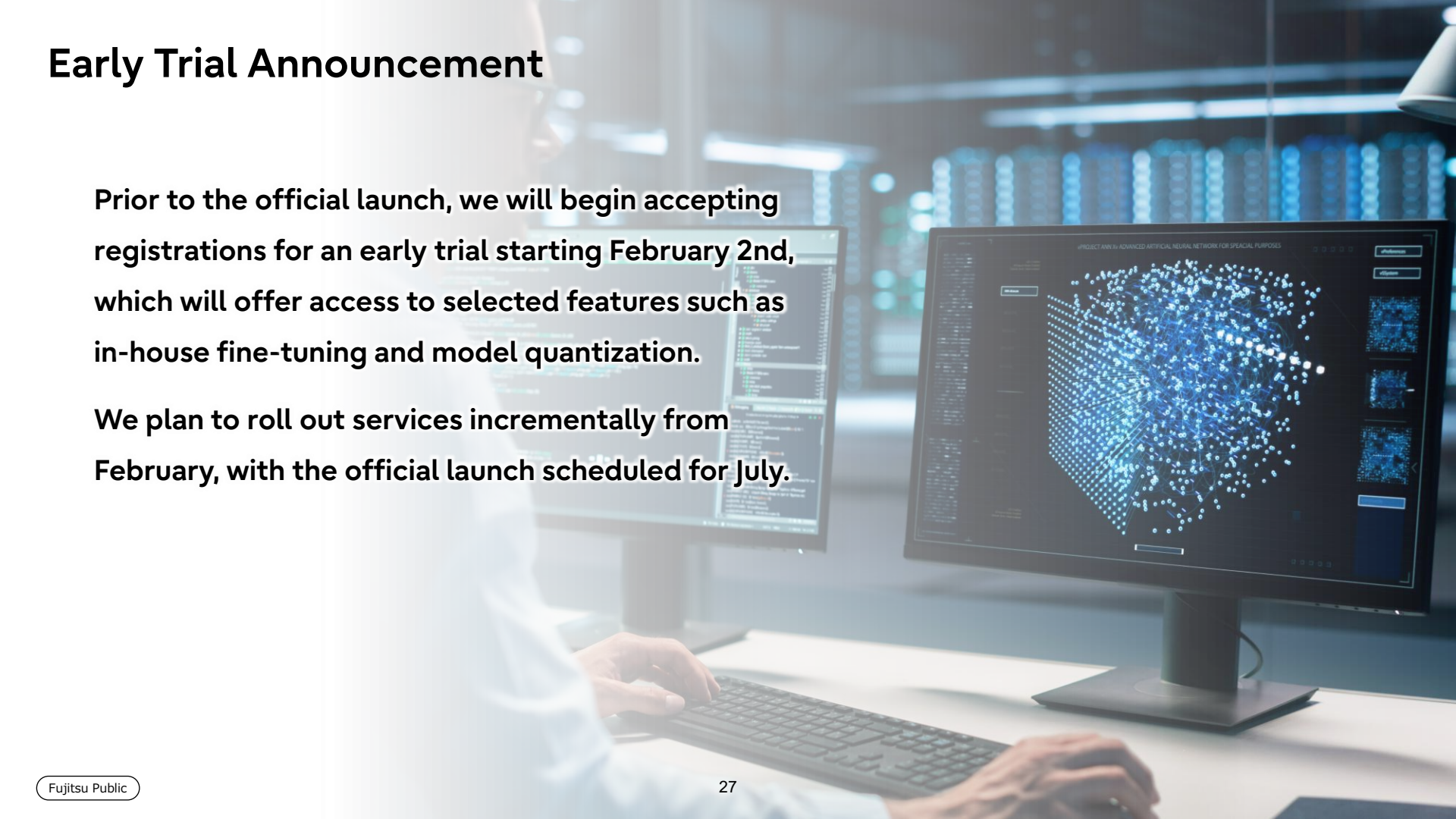
Pre-release trial

© 2026 Fujitsu Limited

Early Trial Announcement

Prior to the official launch, we will begin accepting registrations for an early trial starting February 2nd, which will offer access to selected features such as in-house fine-tuning and model quantization.

We plan to roll out services incrementally from February, with the official launch scheduled for July.



Trial Specifications

uvance

Name	Fujitsu Kozuchi Enterprise AI Factory Early Trial
Trial Overview	A free trial environment is available, allowing users to experience a limited-feature version of Fujitsu Kozuchi Enterprise AI Factory, scheduled for official release in July 2026.
Target Audience	Customers, Fujitsu Group Companies
Trial Environment and Features	・Private AI Platform on PRIMERGY ・Fujitsu Generative AI for Cohere (Takane) ・In-house fine-tuning functionality ・Vulnerability Scanner/Guardrail Function ・LLM Quantization Function (*1) ・AI Agent Framework(*2)
Pricing	Free of charge for environment usage (service support, etc., is generally chargeable).
Support	Email-based Q&A
Usage Period	The usage period will be determined during the preliminary consultation. (Estimated: 2 to 4 weeks)
Application Start	February 2, 2026
Notes	fter conducting preliminary interviews and assessments, we will select the appropriate trial environment and features and determine their availability. As the number of available trial environments is limited, we may not be able to accommodate all scheduling requests.

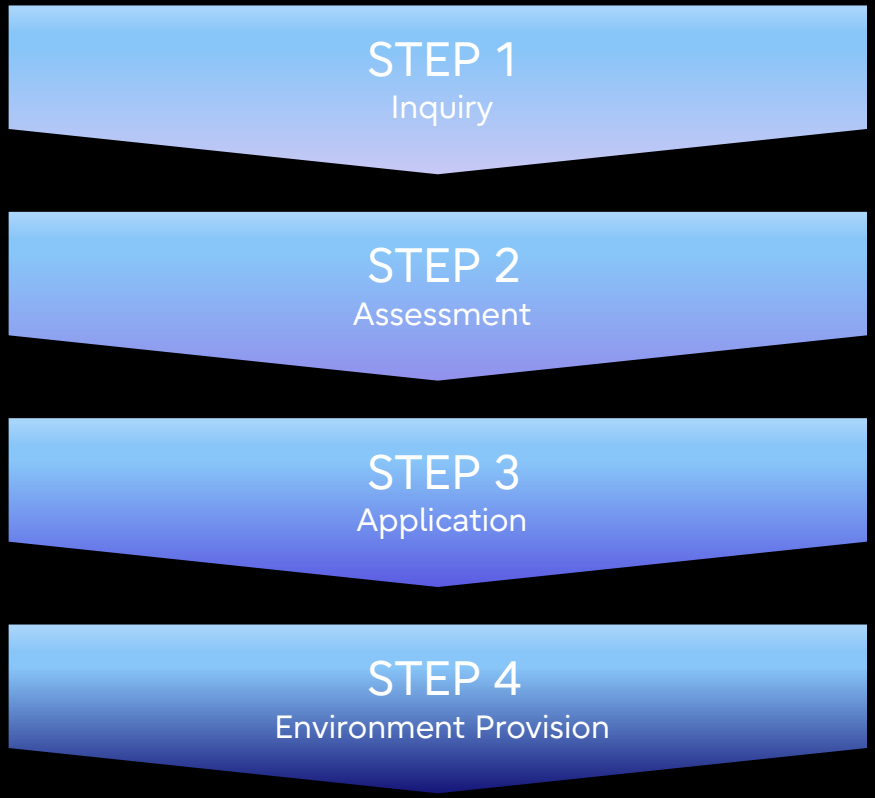
(*1) The service provider will perform the quantization of Takane and provide it. The functionality allowing users to quantize LLMs themselves will be added in phases.

(*2)Dify, standard on the Private AI Platform on PRIMERGY, is available for use. AI Agent functionality will be added in phases.

© 2026 Fujitsu Limited

Steps to Start Your Trial

uvance



- ✓ For inquiries, please contact the trial desk through your assigned sales representative.
- ✓ We will conduct interviews regarding your planned evaluation content, features to be used, usage schedule, and prospects for future full-scale utilization.
- ✓ Based on the interview results, we will determine the feasibility of the trial and adjust the schedule.
- ✓ Formal application will proceed based on the assessment outcomes.
- ✓ We will set up and provide the environment.

© 2026 Fujitsu Limited

© 2026 Fujitsu Limited