Fujitsu Research
Strategy Briefing Session

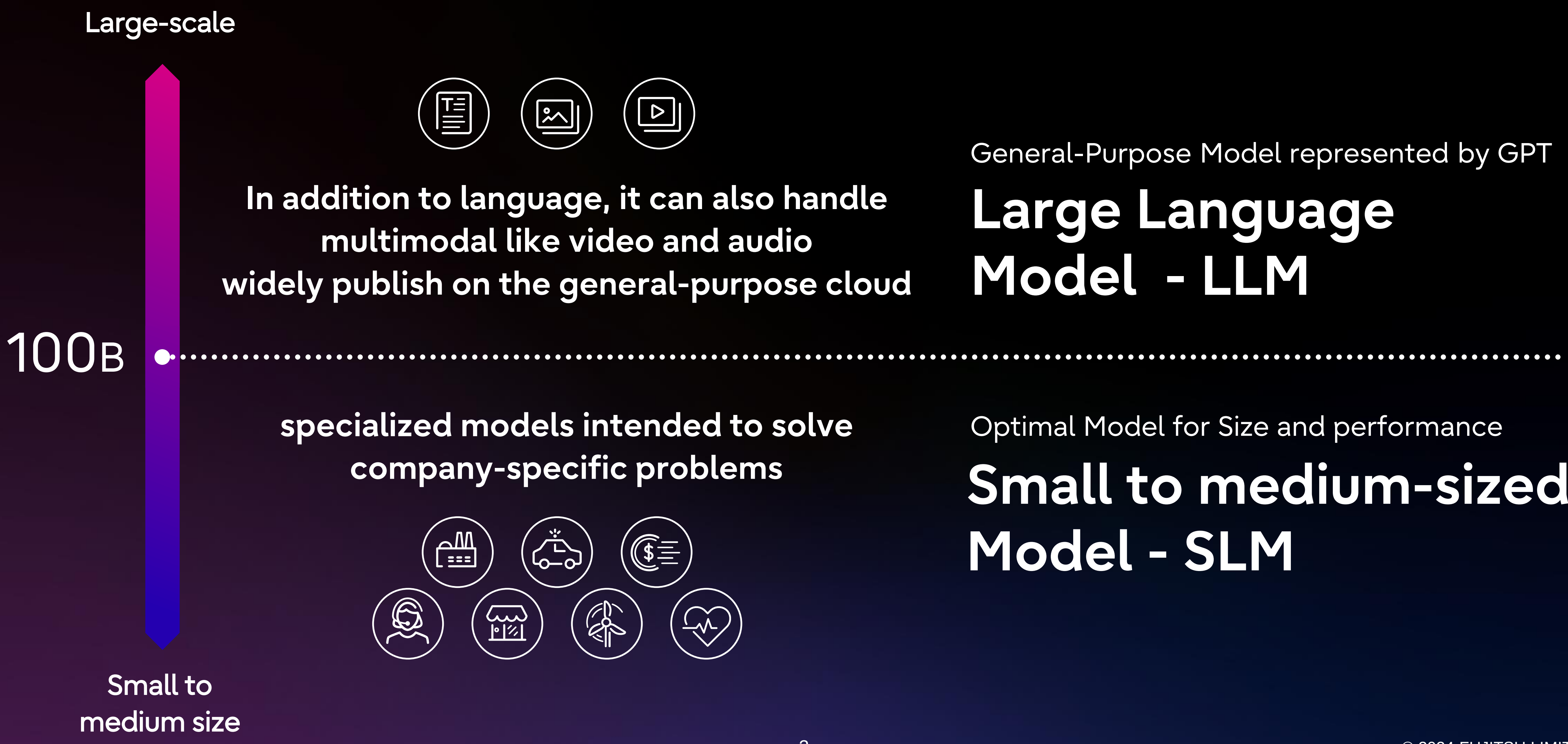
Research Strategies in the field of AI

Generative AI framework for enterprises

June 4, 2024

Toshihiro Sonoda
Head of Artificial Intelligence Laboratory
Fujitsu Research
Fujitsu Limited

Generative AI Market Trends



In addition to language, it can also handle multimodal like video and audio widely publish on the general-purpose cloud

General-Purpose Model represented by GPT

Large Language Model - LLM

specialized models intended to solve company-specific problems

Optimal Model for Size and performance

Small to medium-sized Model - SLM

Fujitsu's Generative AI Strategy

Growing market for generic LLM and small to medium specialized LLM

❯ Focus on specialized LLM for enterprise needs

Three challenges for enterprise use of generative AI

1. Can't handle the variety and volume of data a company has
2. Inability to quickly generate LLMs specific to business know-how and processes
3. Difficulty in complying with corporate rules and regulations

Solve the challenges of using generative AI in enterprises and eliminate security concerns

Generative AI framework for enterprises

Aiming to become a global top player supporting the use of generative AI in enterprises

Fujitsu's Initiatives in Generative AI for Enterprises

Creating an environment where 124,000 global employees can utilize generative AI and implementing it internally

Publishing Conversational Generative AI for enterprises on “Fujitsu Kozuchi”

Developed a large language model, “Fugaku-LLM”, trained on supercomputer “Fugaku”

- Trained a 13 billion parameter model from scratch with proprietary data
- Fujitsu is responsible for speeding up computations and communications, as well as pre-training and subsequent fine-tuning

Generative AI framework for enterprises

FUJITSU

Solving the challenges of using Generative AI in the enterprise and addressing security concerns

Three technologies that make up Fujitsu's generative AI framework for enterprises

World's Top Performance

Knowledge Graph Extended RAG

Solve the handling of large-scale and diverse enterpriase data

10 million characters support

World's Top-Level performance

Generative AI Amalgamation Technology

Flexibly customization for Gen AI to adapt to changing the corporate needs

0 Customization

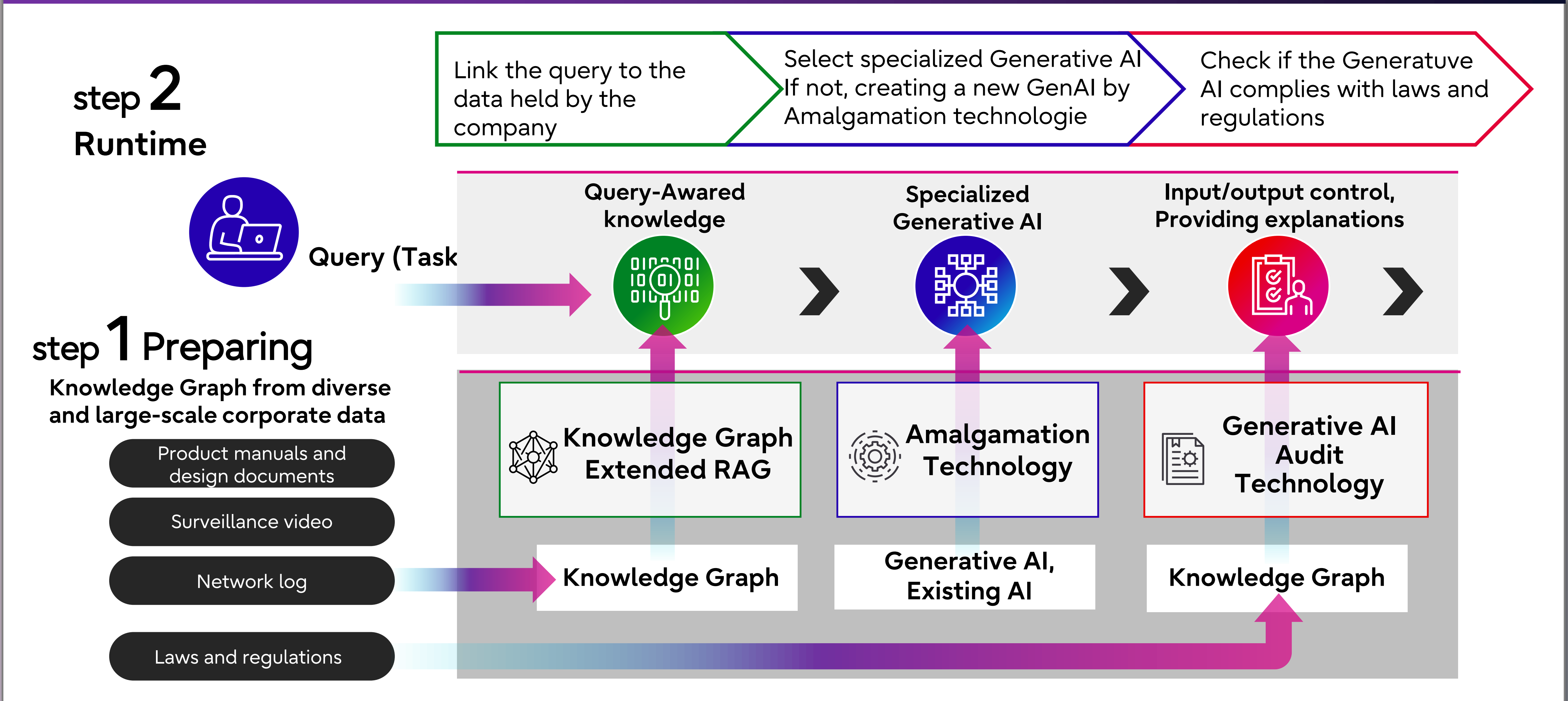
World's first

Generative AI Audit Technology

Eliminate concerns about using Gen AI by controlling the behavior of Generative AI

Compliance with laws & regulations

How generative AI framework for enterprises works



step 2
Runtime

Link the query to the data held by the company

Select specialized Generative AI If not, creating a new GenAI by Amalgamation technologie

Check if the Generatuve AI complies with laws and regulations

Query (Task

Query-Awared knowledge

Specialized Generative AI

Input/output control, Providing explanations

step 1 Preparing

Knowledge Graph from diverse and large-scale corporate data

Product manuals and design documents

Surveillance video

Network log

Laws and regulations

Knowledge Graph Extended RAG

Amalgamation Technology

Generative AI Audit Technology

Knowledge Graph

Generative AI, Existing AI

Knowledge Graph

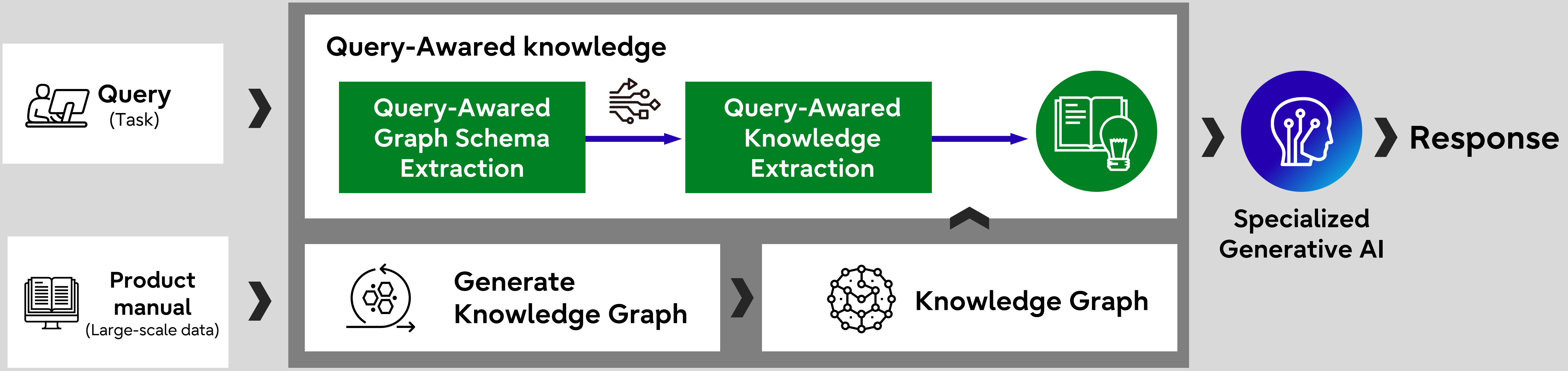
RAG: Retrieval Augmented Generation How LLM captures and uses unlearned knowledge/Knowledge Graph: A network of knowledge that systematically represents various kinds of knowledge in a graph structure

Knowledge Graph Extended RAG

Fujitsu uvance

Analyzing over 10 million characters of entire documents with high precision

- Sequentially process large-scale data handled by companies, such as product manuals, generate a knowledge graph and process large-scale data efficiently
- Extract necessary information from the knowledge graph according to the query – Auxiliary Generative AI inference feature



Query (Task)

Product manual (Large-scale data)

Query-Awared knowledge

Query-Awared Graph Schema Extraction

Query-Awared Knowledge Extraction

Generate Knowledge Graph

Knowledge Graph

Specialized Generative AI

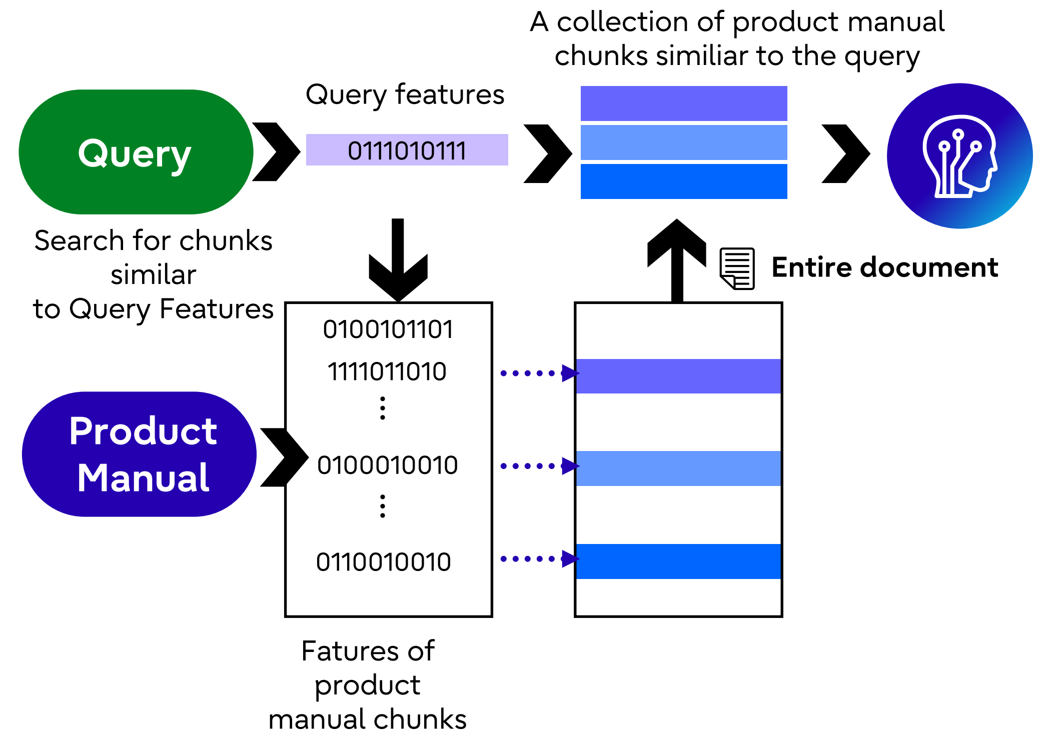
Response

Results Achieved first place in the world in the 'HotpotQA' benchmark that measures the accuracy of complex question answering

Effect of Knowledge Graph Extended RAG

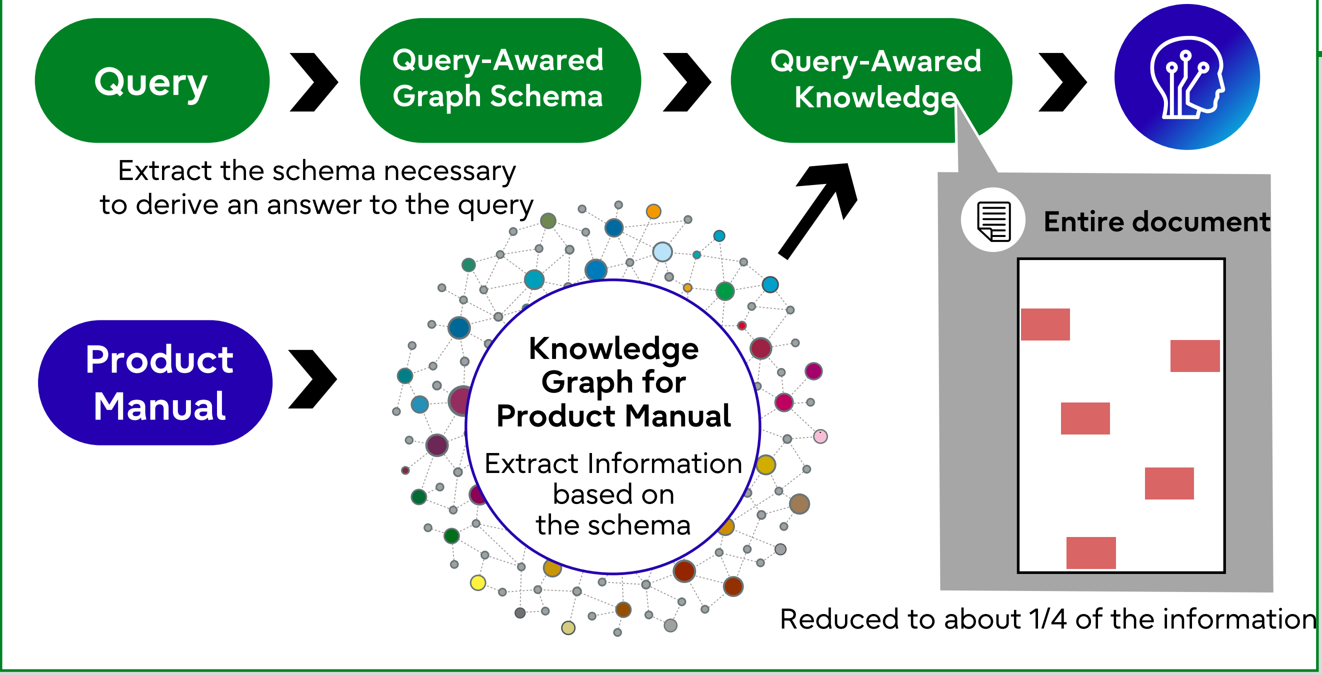
Capable of generating high-precision responses compared to conventional RAG technology

Conventional RAG


A collection of product manual chunks similiar to the query
Query features
Query
0111010111
Search for chunks similar to Query Features
Entire document
Product Manual
0100101101
1111011010
⋮
0100010010
⋮
0110010010
Fatures of product manual chunks

Knowledge Graph Extended RAG

Compared to conventional RAG, it extracts only the information necessary for the answer, reducing the amount of information provided to the generative AI to about 1/4 of the conventional RAG, and achieved first place in the 'HotpotQA' benchmark


Query
Query-Awared Graph Schema
Query-Awared Knowledge
Extract the schema necessary to derive an answer to the query
Product Manual
Knowledge Graph for Product Manual
Extract Information based on the schema
Entire document
Reduced to about 1/4 of the information

Applications and Effect of Knowledge Graph Extended RAG

DEMO Exhibition

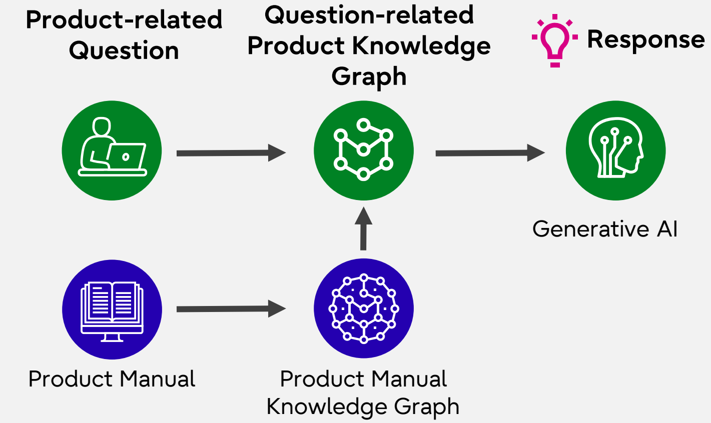
10 million characters support

Product Manual Q & A

Product manuals cannot provide answers that overlook the entire content of over 10 million characters

Confirmed effectiveness in Q&A for product manual with over 10 million characters

Properly integrate information across multiple pages to generate the best answer



Product-related Question

Question-related Product Knowledge Graph

Response

Generative AI

Product Manual

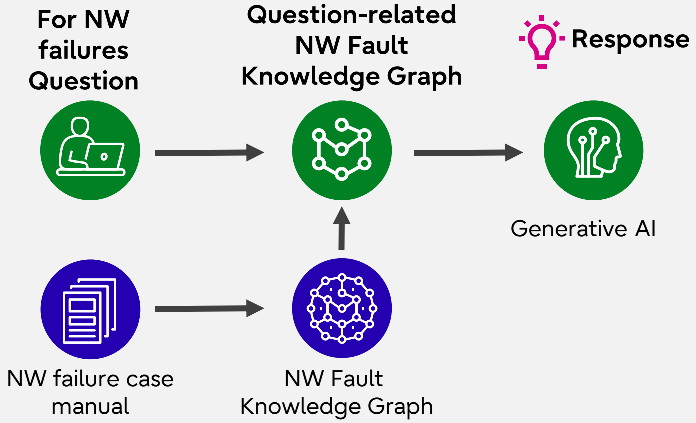
Product Manual Knowledge Graph

Network Log Analysis

Unable to identify the cause of network failures from massive network logs and past failure cases

Applied to mobile network connection failures, confirmed effectiveness

Generate KG from different failure cases and streamline failure recovery by listing potential causes



For NW failures Question

Question-related NW Fault Knowledge Graph

Response

Generative AI

NW failure case manual

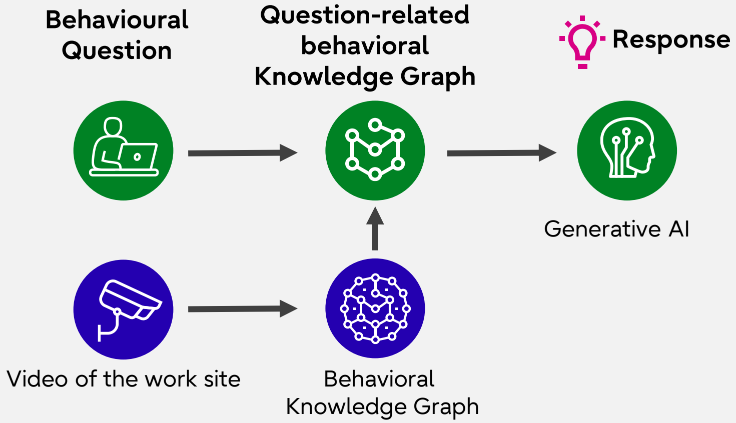
NW Fault Knowledge Graph

Work Analysis through Video

Unable to handle aggregation and statistical information of large amounts of video data over a long period of time

It is possible to check the long-term situation of workers from the video of the work site

Testing in the actual warehouse



Behavioural Question

Question-related behavioral Knowledge Graph

Response

Generative AI

Video of the work site

Behavioral Knowledge Graph

Amalgamation Technology

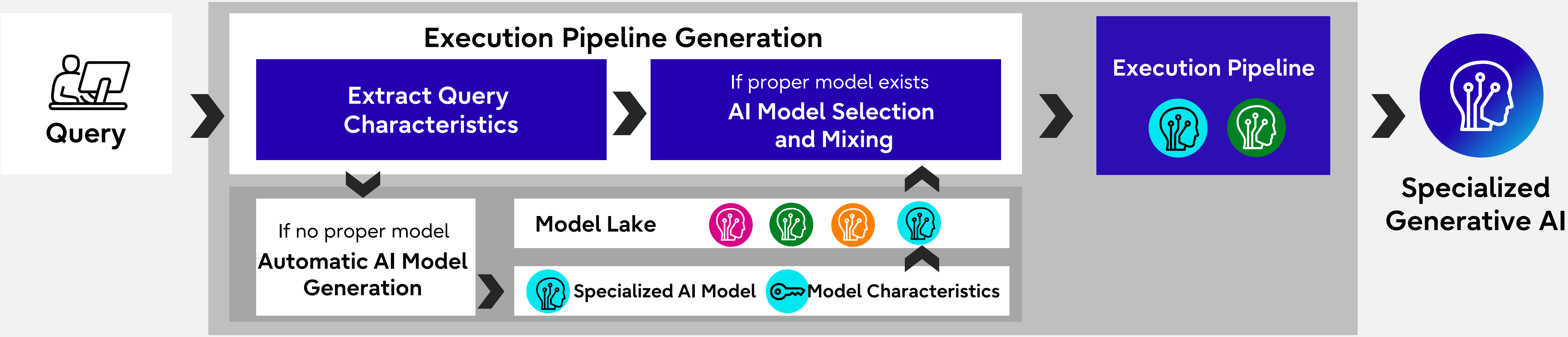
Automatically generate highly effective specialized generative AI easily, without the need for customer customization such as prompt engineering or fine-tuning

- **Select Specialized Generative AI**

  Select the AI model required to perform the task from Query Characteristics*[1] and Model Charactarisitcs*[2]

- **Automatic Generation of Specialized Generative AI**

  Automatic generation of the required AI model if the appropriate AI model is not available


Query
Execution Pipeline Generation
Extract Query Characteristics
If proper model exists
AI Model Selection and Mixing
If no proper model
Automatic AI Model Generation
Model Lake
Specialized AI Model
Model Characteristics
Execution Pipeline
Specialized Generative AI

Results

Same as GPT-4V in video detection, highest performance achieved in Japanese open model
Check the effectiveness of company-specific tasks such as contract compliance checks and support operations efficiency

*1 Query Charactaristics: Indicators representing the characteristics of the user's query used to select models to handle this task
*2 Model Charactaristics: Characteristics of the enterprise information added when generating specialized models. Used for conformance checking as a model to process queries from users

Application and effects of Amalgamation Technology

DEMO Exhibition

0 Customization

Contract Compliance Check

It takes a tremendous amount of time to check software contracts and usage status

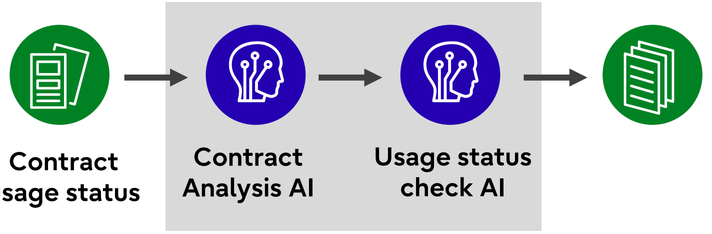
Match with the contractcontents

Complementing with operational knowledge

There is a need to teach Fujitsu's verification know-how to the generative AI

Autogenerate specialized generative AI without having to spend months in prompt engineering

30% man-hour reduction


Contract usage status → Contract Analysis AI → Usage status check AI →

Streamlined Support Operations

The assignment of incident responders is laity, which causes delays

Contact Skills

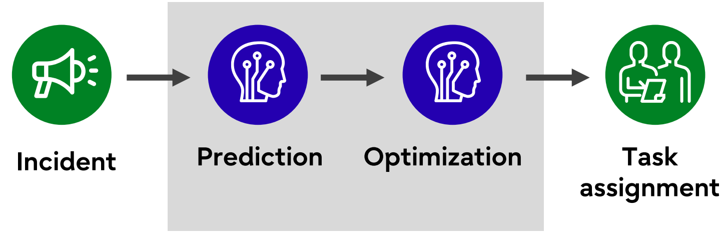
SLA compliance

Urgency

Experts are needed for predictive optimization model development

Automatically generate specialized generative AI to solve complex task assignments

Work efficiency 25% improvement


Incident → Prediction → Optimization → Task assignment

Optimal Driver Assignment

"2024 Issue" of logistics Shortage of 200,000 drivers

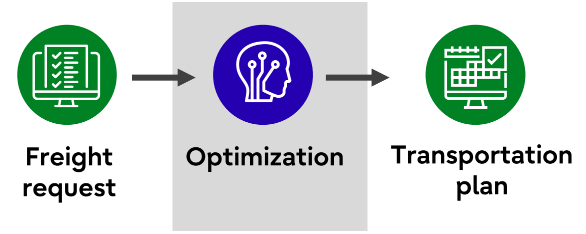
Compliance with laws & regulations

Shortage of manpower

Experts are needed for optimization formalization

Automatically generate specialized generative AI that can immediately execute formalization that takes experts several weeks

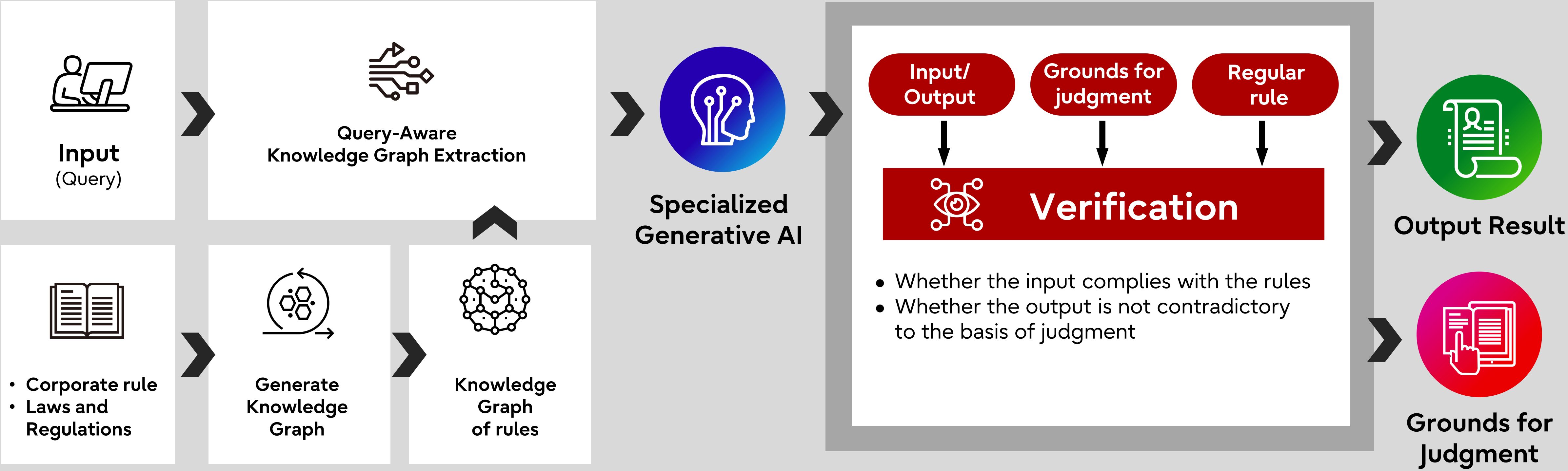
Planning time 95% Reduction


Freight request → Optimization → Transportation plan

Generative AI Audit Technology

Control the behavior of generative AI with a knowledge graph, complying with corporate rules and laws

- Utilize Knowledge Graph that corresponds to laws and corporate rules **to verify compliance with input rules**.
- By analyzing the basis on which the generative AI derived its output, we provide explanability for the grounds of judgment and determine hallucination



- Whether the input complies with the rules
- Whether the output is not contradictory to the basis of judgment

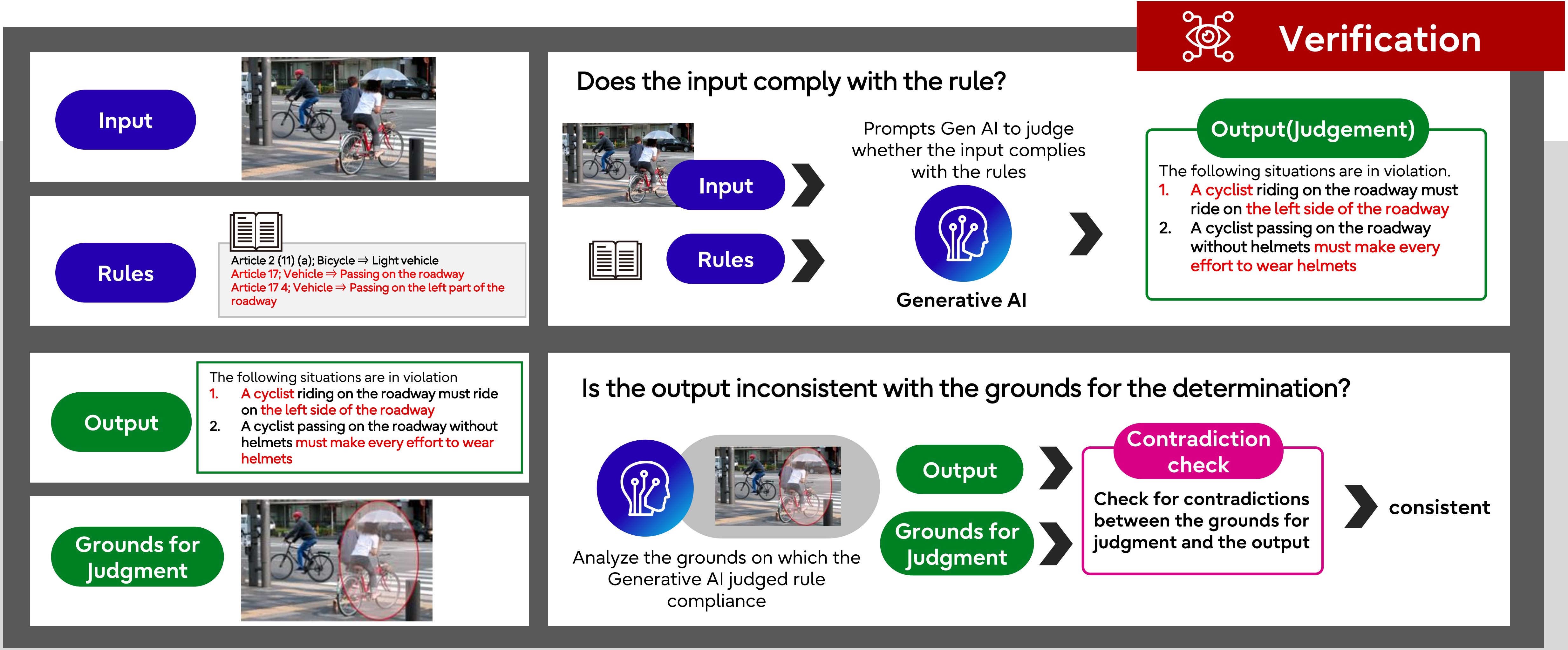
Generative AI Audit Technology

DEMO Exhibition

Compliance with laws & regulations

FUJITSU

Rule compliance verification and providing an explanation of the output basis



Verification

Input

Rules

Article 2 (11) (a); Bicycle ⇒ Light vehicle
Article 17; Vehicle ⇒ Passing on the roadway
Article 17 4; Vehicle ⇒ Passing on the left part of the roadway

Does the input comply with the rule?

Input

Rules

Prompts Gen AI to judge whether the input complies with the rules

Generative AI

Output(Judgement)

The following situations are in violation.

1. A cyclist riding on the roadway must ride on the left side of the roadway
2. A cyclist passing on the roadway without helmets must make every effort to wear helmets

Output

The following situations are in violation

1. A cyclist riding on the roadway must ride on the left side of the roadway
2. A cyclist passing on the roadway without helmets must make every effort to wear helmets

Grounds for Judgment

Is the output inconsistent with the grounds for the determination?

Analyze the grounds on which the Generative AI judged rule compliance

Output

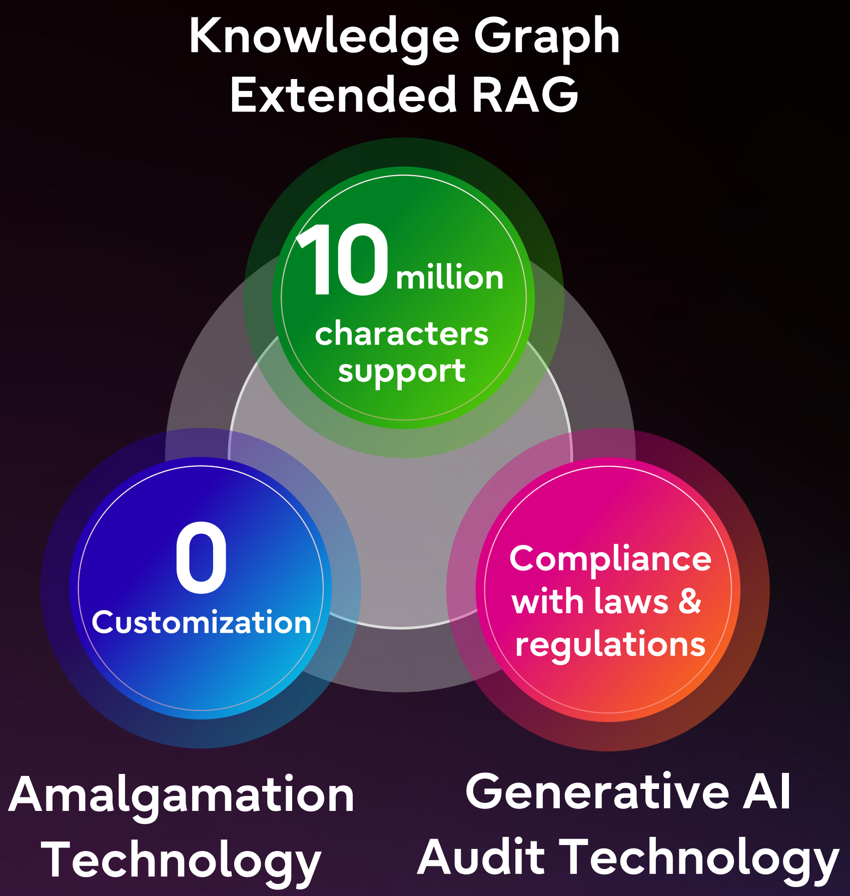
Grounds for Judgment

Contradiction check

Check for contradictions between the grounds for judgment and the output

consistent

FUJITSU


Knowledge Graph Extended RAG
10 million characters support
0 Customization
Amalgamation Technology
Compliance with laws & regulations
Generative AI Audit Technology

By Fujitsu Kozuchi

We plan to gradually release Fujitsu Enterprise Generative AI Framework from July

Fujitsu aims to be a global top player leading the utilization of Generative AI in enterprises

Fujitsu Research
Strategy Briefing Session

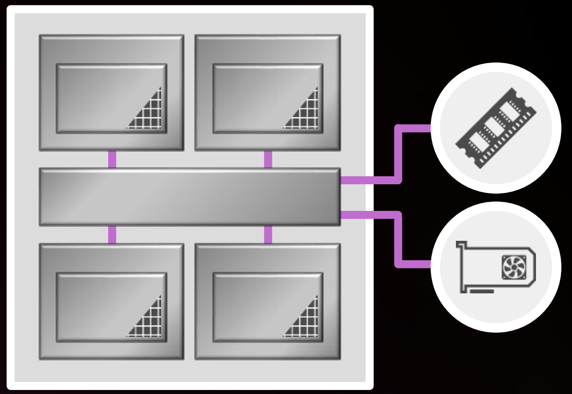
Processor for next generation data centers

FUJITSU-MONAKA

June 4, 2024

Naoki Shinjo
SVP, Head of Advanced Technology Development Unit
Fujitsu Research
Fujitsu Limited

FUJITSU-MONAKA



- Armv9-A Architecture
- Arm SVE2 for AI and HPC
- 3D chiplet
  - Core die 2nm
  - SRAM die/IO die 5nm
- 144 cores x 2 sockets (288 cores per node)
- Ultra low voltage for energy-efficiency
- Confidential Computing for security
- DDR5 12 channels
- PCI Express 6.0 (CXL3.0)
- Air cooling

To be shipped in 2027

* This presentation is based on results obtained from a project subsidized by the New Energy and Industrial Technology Development Organization (NEDO)

FUJITSU

Next-generation high-performance, energy-efficient, Japan-made processor for a carbon neutral digital society

High-speed data processing platform

Achieve high-speed processing of computing workloads, particularly AI workloads (2x faster than competing CPUs)

Balance of energy efficiency and performance

Significantly reduce CO2 emissions and power costs with high energy efficiency (2x more efficient than competing CPUs)

High security & reliability

Stable operation technology cultivated in mainframes and high security for cloud utilization

Easy to use

Utilize the Arm software ecosystem, and collaborative design across services, software, and hardware

Achieved through our proprietary technologies such as self-designed microarchitecture and ultra low-voltage technology

Initiating co-creation with various fields to promote utilization in diverse applications

Data Center

Telecom

Security

Arm-based CPU
FUJITSU-MONAKA

FUJITSU-MONAKA
Software

Meeting the growing AI demands of Data Centers

Pursuing AI performance

2x the performance and energy efficiency compared to competing CPUs

Expanding AI software

Wide range of domain-specific software stacks

FUJITSU

Covering a wide range of software stacks, including AI and HPC

Product Delivery

Customer Use Cases
- Surrogate Models SVR
- LLM Software Applications

Fujitsu Computing as a Service
- Scikit Learn Use Cases
- Hugging Face Use Cases

Fujitsu Kozuchi
- Causal Inference
- Ambient Authentication

Open-Source Contributions

API Microservices Platform (FUJITSU-MONAKA Green HPC API Server) @ FRIPL*

OpenMathLib /OpenBLAS	Math Library, NumPy, OpenMP
UXL foundation	oneAPI Ecosystem, oneDAL &oneDNN
PyTorch	FUJITSU-MONAKA ARMImprovements
Linaro	Kubernetes and OpenStack OSS

Software Delivery

PyPi | Docker | Containers | Reference Implementations | Computing Workload Broker

Continuous Integration and Deployment using MOCHI and Konark Platform

Collaborations

Internal Teams: SW R&D | Math Lib | Platform | AI Solutions | Computing | Compiler

External Organizations: ARM | UXL | MIT | IISc

AI Software Frameworks

Machine Learning: Scikit-Learn, Multithreading, XGBoost, NumPy, Pandas, BLAS

Deep Learning: LLM's, Vision, NLP, Hugging Face, TensorFlow/PyTorch, OpenVINO, oneDNN, Inductor

Big Data Analytics: PostgreSQL, PySpark, VectorDB, Data Intelligence

Data Security: Red Hat, Secured HW/SW, Software Guard Extensions, OpenShift, Confidential Computing

Software Stack Selection

Quantitative Metrics: Downloads | GitHub | Market Adoption | Search Trends

Qualitative Metrics: Arm Enablement | Release Freq. | Innovation Scope | Use Case

Cutting Edge Applications

Healthcare
- ❑ Drug Discovery
- ❑ Gene Prediction

Manufacturing
- ❑ Defect Detection
- ❑ Preventive Maintenance

Retail
- ❑ Recommendation
- ❑ SCM Forecasting

Banking
- ❑ HF Trading
- ❑ Fraud Detection

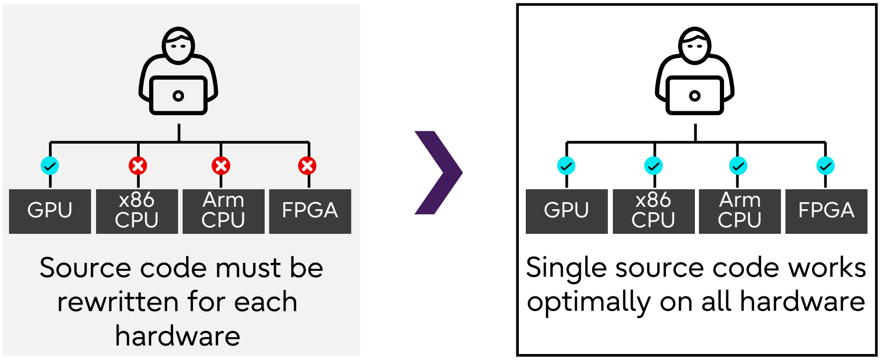
*Fujitsu Research of India Private Limited

Introduction of software technology development examples to reduce barriers to adoption

Development of Unified Acceleration technology to utilize various AI accelerators with a single code

- As a founding member, Fujitsu is actively involved in the UXL Foundation, a consortium of companies promoting the adoption of Unified Acceleration, which aims to enable the use of various CPUs and accelerators with a single source code
- Fujitsu is developing global foundation software to utilize the Arm-based CPU FUJITSU-MONAKA as an AI accelerator
- Aims to create an environment where customers can easily maximize the AI performance of FUJITSU-MONAKA by 2027

Benefits of Unified Acceleration



Source code must be rewritten for each hardware

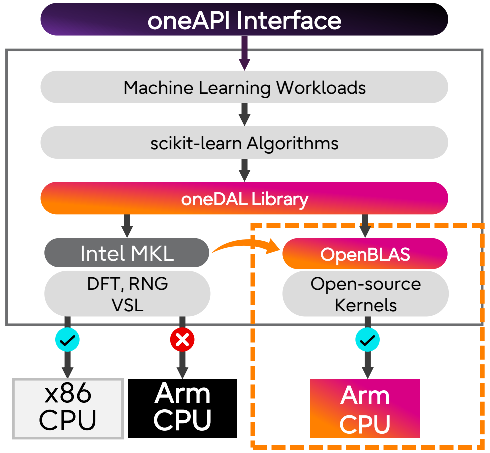
Single source code works optimally on all hardware

High environment migration cost	Reduced environment migration cost
Reduced maintainability	Optimal hardware selection is possible

This technology will expand the use of AI with FUJITSU-MONAKA

Latest example ： First successful Arm enablement for oneDAL

Successfully replaced MKL MATH functions with optimized open-source compute kernels of OpenBLAS

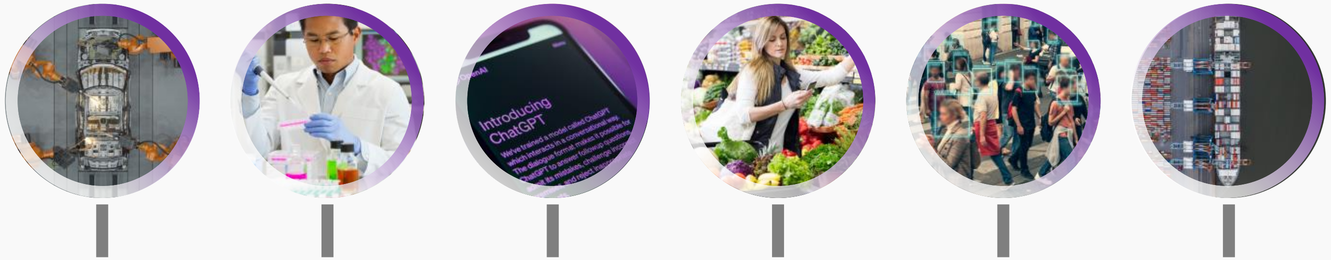


Enables the use of high-speed processing routines in large-scale computations

Expanding Arm enablement to build an AI solution development platform

FUJITSU

FUJITSU-MONAKA will solve customer issues as an AI infrastructure platform that can be utilized in a wide range of fields



Fujitsu Kozuchi

FUJITSU-MONAKA & Tuned AI softwares

Application	Speech Recognition	Surrogate Model	Generative AI			
Library Framework	OpenBLAS	NumPy/SciPy	scikit-learn	oneDNN	oneDAL	PyTorch/TensorFlow
Middleware OS	Linux	Kubernetes	OpenStack	GCC/LLVM	Confidential Containers	
Firmware Hardware	Many Core	High-Capacity Memory	Low Power	Support AI instructions		

Fujitsu Research
Strategy Briefing Session

Fujitsu's Research Strategy:

Creating New Value by Combining Technology areas

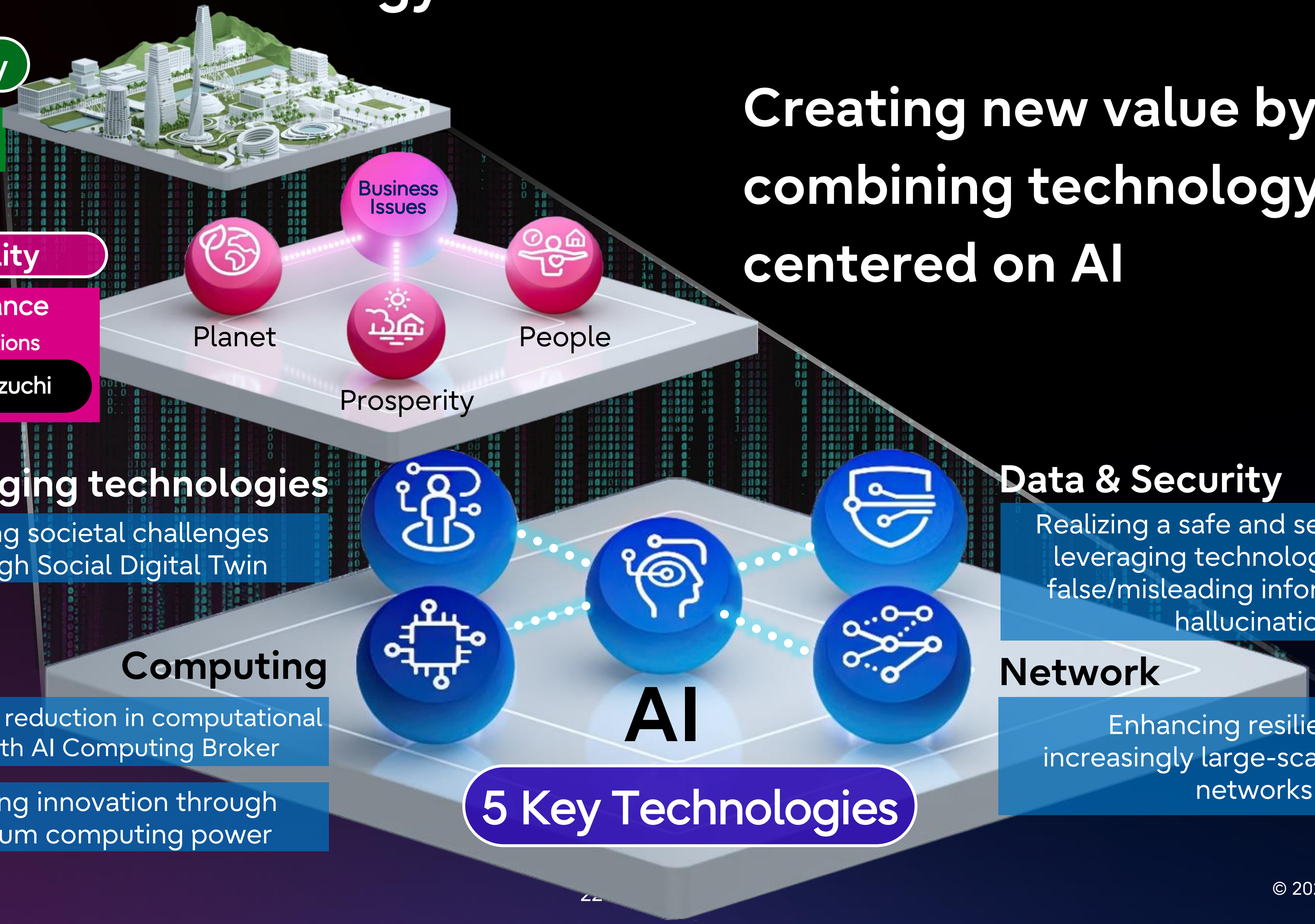
June 4, 2024

Seishi Okamoto
Corporate Executive Officer, EVP
Head of Fujitsu Research
Fujitsu Limited

Fujitsu's Research Strategy

FUJITSU

Creating new value by combining technology areas centered on AI



Sustainable society

Transforming into regenerative enterprise

Materiality

Fujitsu Uvance

Service solutions

Fujitsu Kozuchi

Converging technologies

Solving societal challenges through Social Digital Twin

Computing

Significant reduction in computational cost with AI Computing Broker

Amazing innovation through quantum computing power

Data & Security

Realizing a safe and secure society leveraging technologies against false/misleading information and hallucination

Network

Enhancing resiliency of increasingly large-scale, complex networks

5 Key Technologies

AI × Computing

In 2030,

10% of all electricity generated in the world will be consumed at datacenters

Development of AI will directly affect the global electricity problem

AI Computing Broker

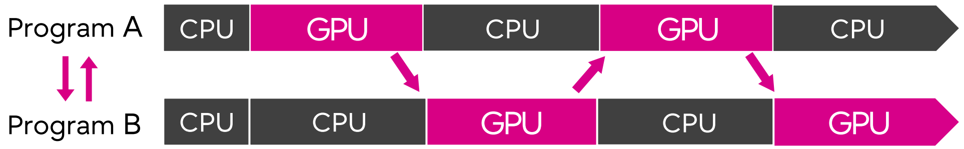
DEMO target New technology

FUJITSU

Drastically reducing power consumption at datacenters

Technology to fully utilize GPUs (up to 100%)

Analyzing the jobs requiring AI calculation using GPU in advance and dynamically allocate those jobs during operation



GPU usage rate of TSUBAME is about 30%

Reducing power consumption by halving the number of GPUs

Enable to reduce power consumption by 10TWh per year by reducing resources requiring AI calculation

Equivalent to annual electricity consumption by about 24 million households in Japan

AI × Data & Security

False and misleading information by AI is the biggest global risk

2024 World Economic Forum

Disinformation and misinformation from generative AI and synthetic content is posing unprecedented social risks by influencing election processes, stock markets, etc.

Addressing a New Societal Challenge

Rulemaking and development of anti-disinformation technologies

Participating in discussions for international governance formation and presenting our proposals

OECD

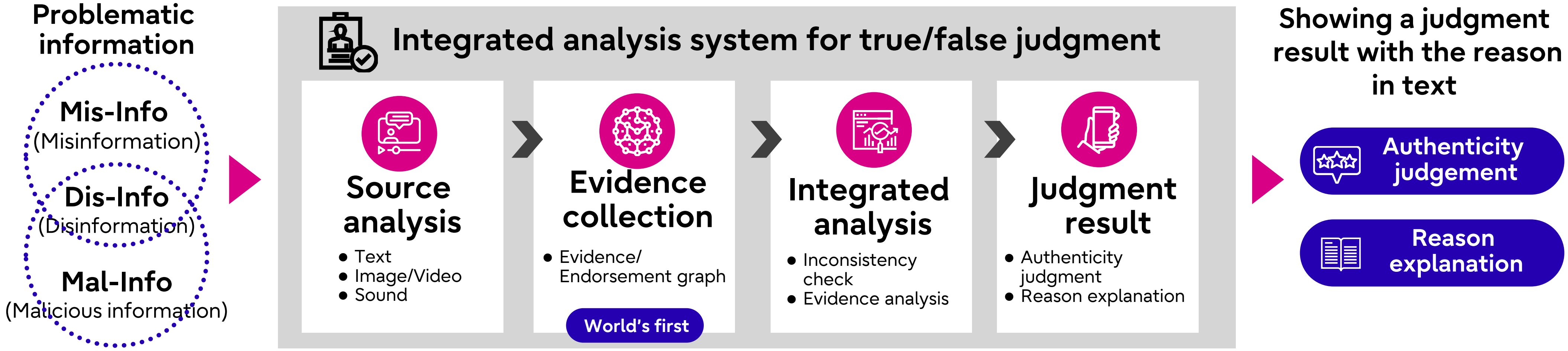
G7 Hiroshima AI Process

MIC*/METI** AI Guidelines for Business

*Ministry of Internal Affairs and Communications/**Ministry of Economy, Trade and Industry

World's first integrated analysis system for authenticity judgment

New technology



Problematic information

- Mis-Info (Misinformation)
- Dis-Info (Disinformation)
- Mal-Info (Malicious information)

Integrated analysis system for true/false judgment

Source analysis
- Text
- Image/Video
- Sound

Evidence collection
- Evidence/ Endorsement graph

World's first

Integrated analysis
- Inconsistency check
- Evidence analysis

Judgment result
- Authenticity judgment
- Reason explanation

Showing a judgment result with the reason in text

Authenticity judgement

Reason explanation

AI × Quantum Computer

Revolutionize the world of AI with exponentially fast quantum computing power

- Large-scale multi-agent AI
- Ultra-personalized AI
- Ultra-low power consumption edge AI

World's First Quantum Machine Learning Technology

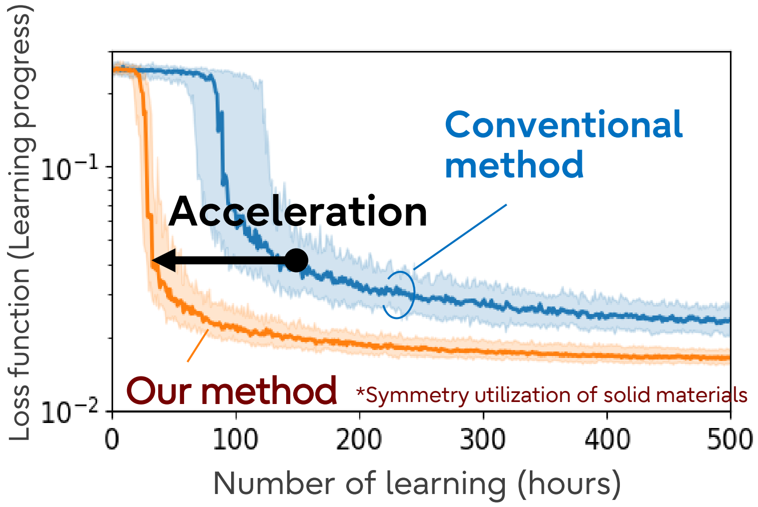
Starting to use the hybrid quantum platform

New technology

World's fastest quantum CNN technology

Predict the properties of solid substances

Learning about the properties of magnetic materials

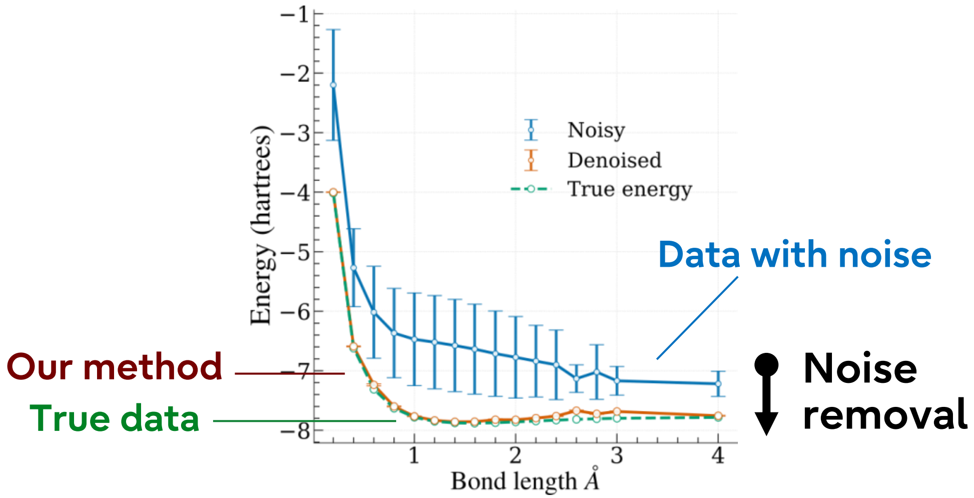


CNN: Convolutional neural network

World's first quantum noise removal technology

Successful data recovery using quantum autoencoder

Energy calculation of lithium hydride



Creating New Value
by Combining Technology Areas Centered on AI