March 15, 2024

SB Intuitions

SB Intuitions' Challenge:
From Present to Future of Japanese LLMs

Board Director & CTO, SB Intuitions Corp.

Daiki Orihara

Daiki Orihara

SB Intuitions Corp.
Board Director & CTO

SoftBank Corp.
Vice President, Data Platform Strategy Division

Leading the development of platforms for application services, the development of unified communications, voice services and security services. Currently managing the research development of a cross integrated platform and the realization of innovative business models using IoT products.
Also CTO & Engineering Division Head for SB Intuitions Corp. and member of the Information Technology Federation of Japan.

SB Intuitions｜Promoting GenAI development

SoftBank

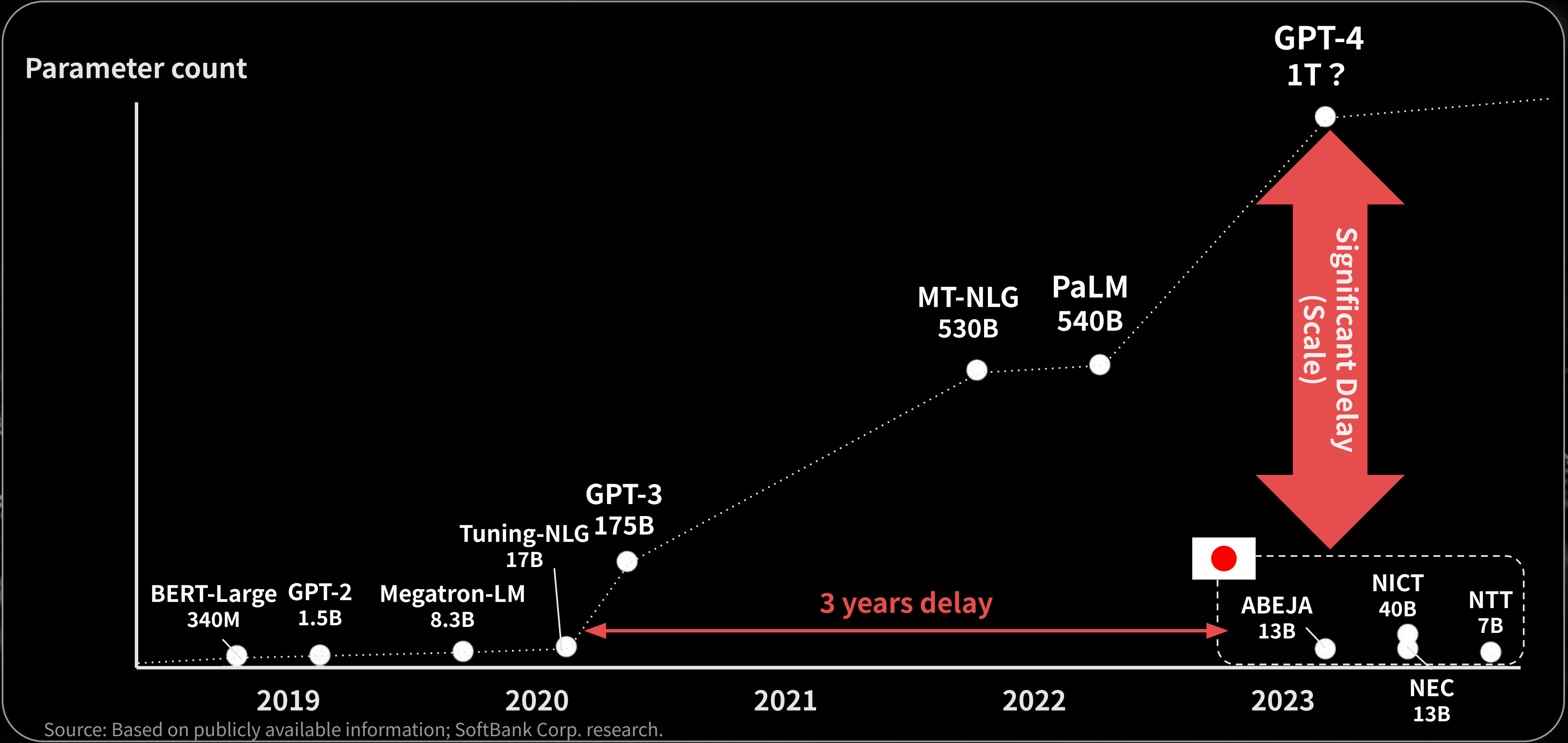
Investment(100%)

SB Intuitions

- Research and develop Japanese LLMs
- Provide safe & secure AI services tailored to Japanese culture and business practices

Gathered experienced engineers in GenAI from group companies

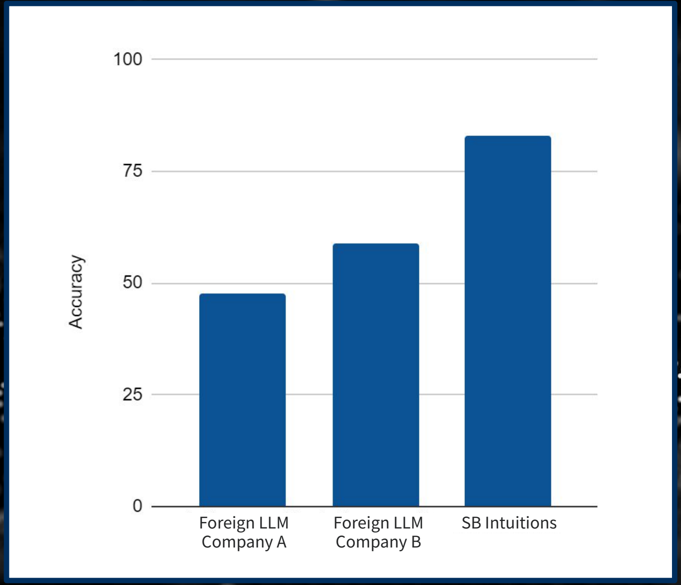
Japanese LLMs are falling behind global counterparts



Source: Based on publicly available information; SoftBank Corp. research.

The difference in Japanese proficiency between Japanese-made models and globally recognized models

The evaluation benchmark "AI-Oh(AI王)" includes many questions that delve deep into Japanese culture and history, as well as complex knowledge. It has been found that the model being developed at SB Intuitions has already surpassed the performance of foreign models



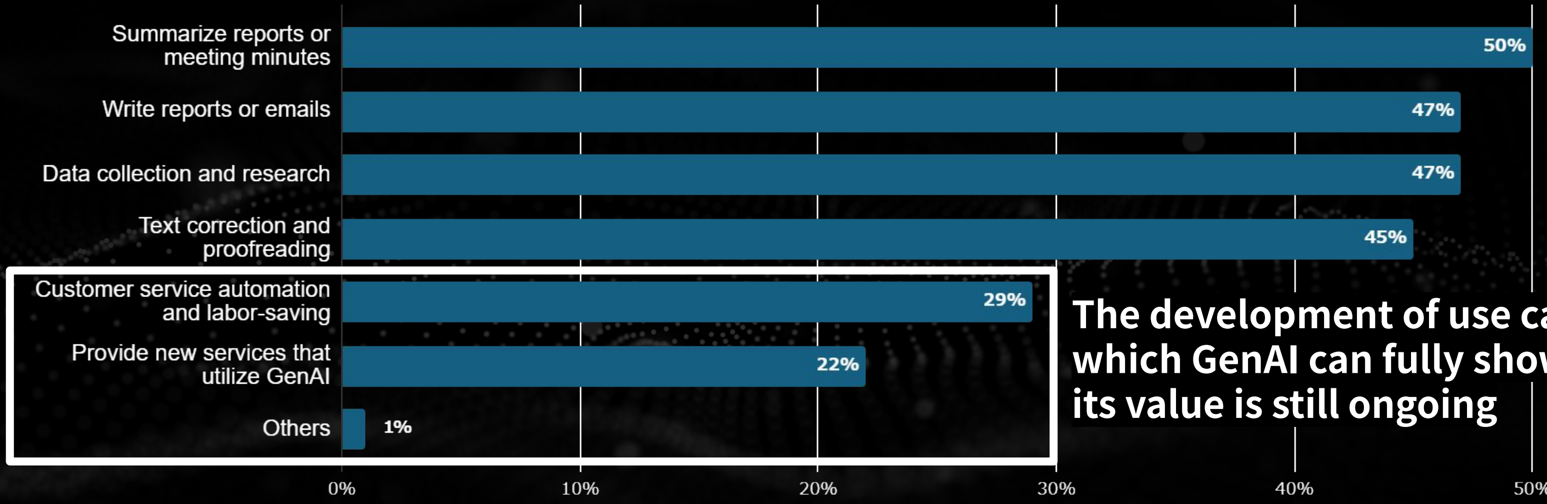
Results based on "AI-Oh(AI王)" benchmark
※ Evaluation conducted by SB Intuitions
※ Evaluation of foreign LLMs based on released models/API publicly available on February 2024

Question Example (Translated from Japanese)	Answer
What is the light brown color that serves as a guide for the ideal caramelization when sautéing onions?	飴色 (AMEIRO)
What is the term for a deposit account with a bank or financial institution that has been dormant for an extended period of time, with no deposits or withdrawals?	休眠口座 (KYU-MIN-KOUZA)
Using a single letter of the alphabet, what is the term for the competitive activity in which multiple players compete against each other in a computer game?	eスポーツ (e-sports)

Use cases of GenAI in the Japanese market

Although simple use cases such as summarization and proofreading are already growing, there are still ample opportunities for further integrations of GenAI into various systems

Q: Provide the applicable use cases of Gen AI that are currently under consideration or have already been realized. (n=799)

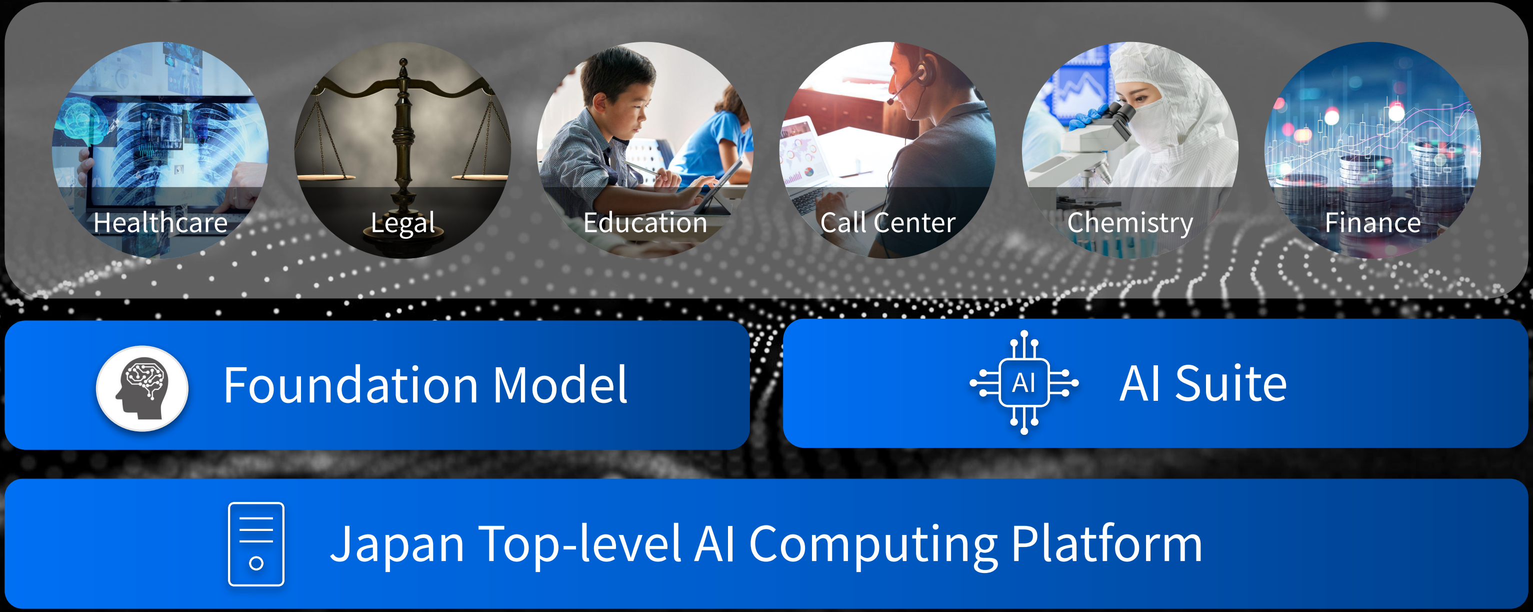


The development of use cases in which GenAI can fully showcase its value is still ongoing

Source: Survey on the Current State of Generative AI, Autumn 2023 | PwC
https://www.pwc.com/jp/ja/knowledge/thoughtleadership/2023/assets/pdf/generative-ai-survey2023_autumn.pdf

SB Intuitions’ Vision

SB Intuitions' GenAI serves as a competitive advantage across all industries, fueling growth and progress in society



Healthcare

Legal

Education

Call Center

Chemistry

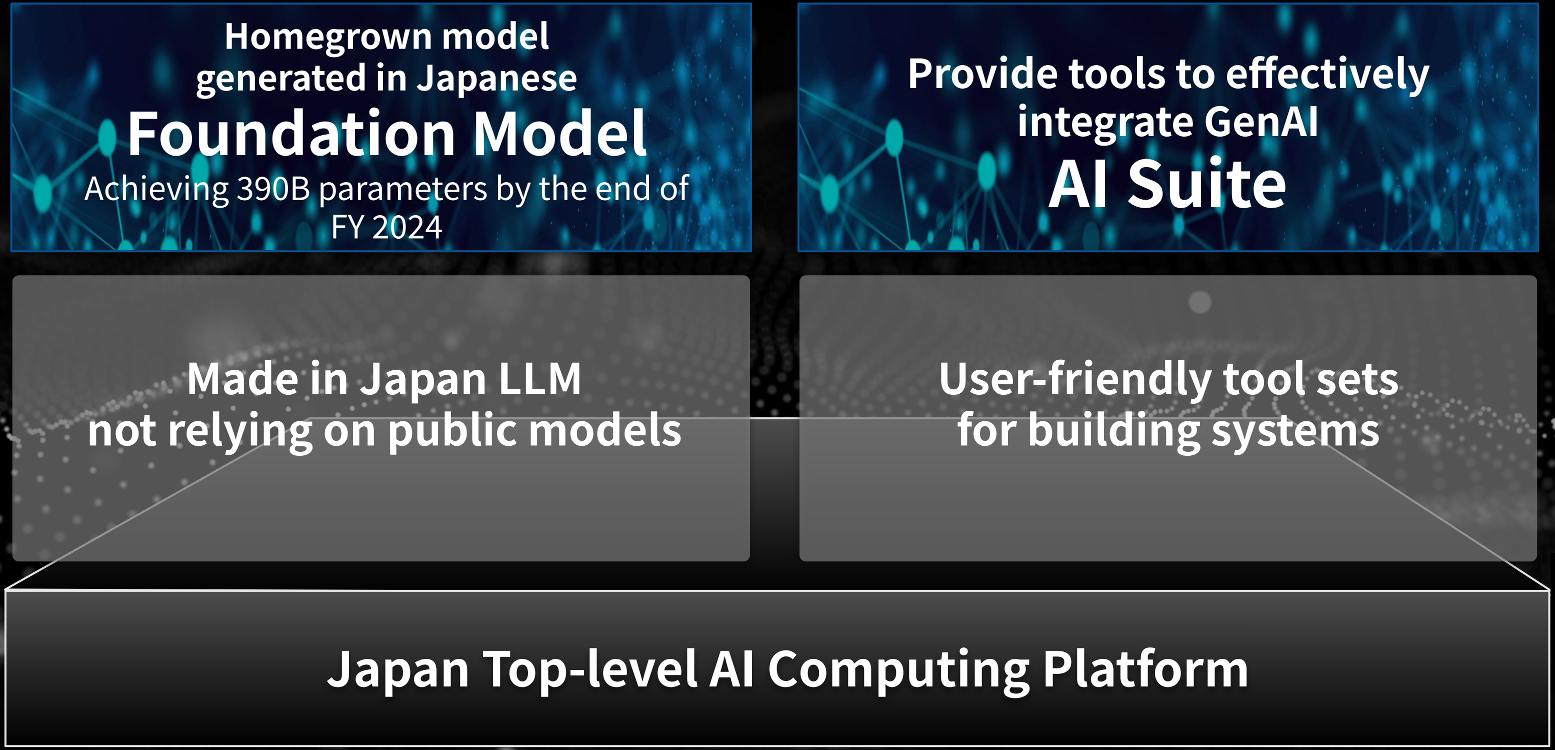
Finance

Foundation Model

AI Suite

Japan Top-level AI Computing Platform

SB Intuitions' Foundation Model and AI Suite



Homegrown model generated in Japanese

Foundation Model

Achieving 390B parameters by the end of FY 2024

Provide tools to effectively integrate GenAI

AI Suite

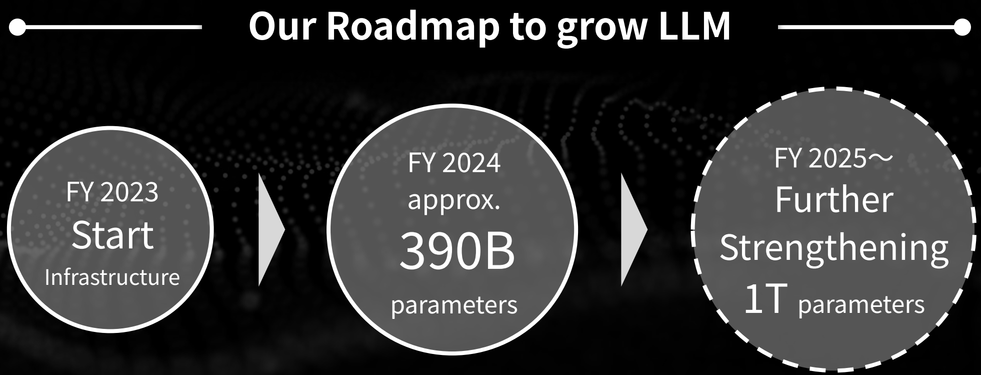
Made in Japan LLM not relying on public models

User-friendly tool sets for building systems

Japan Top-level AI Computing Platform

SoftBank’s Japan Top-level Generative AI computing platform

- Japan’s largest computing platform for GenAI development* started operating in October 2023
- Comprises an NVIDIA DGX SuperPOD™ AI supercomputer with over 2,000 NVIDIA Tensor Core GPUs, NVIDIA Networking and NVIDIA AI Enterprise software
- While using the computing platform in phases, SoftBank and SB Intuitions plan to complete all investments and construction within FY 2023 (ending March 31st 2024), and quickly start providing to universities, research institutions, companies and other entities


Our Roadmap to grow LLM

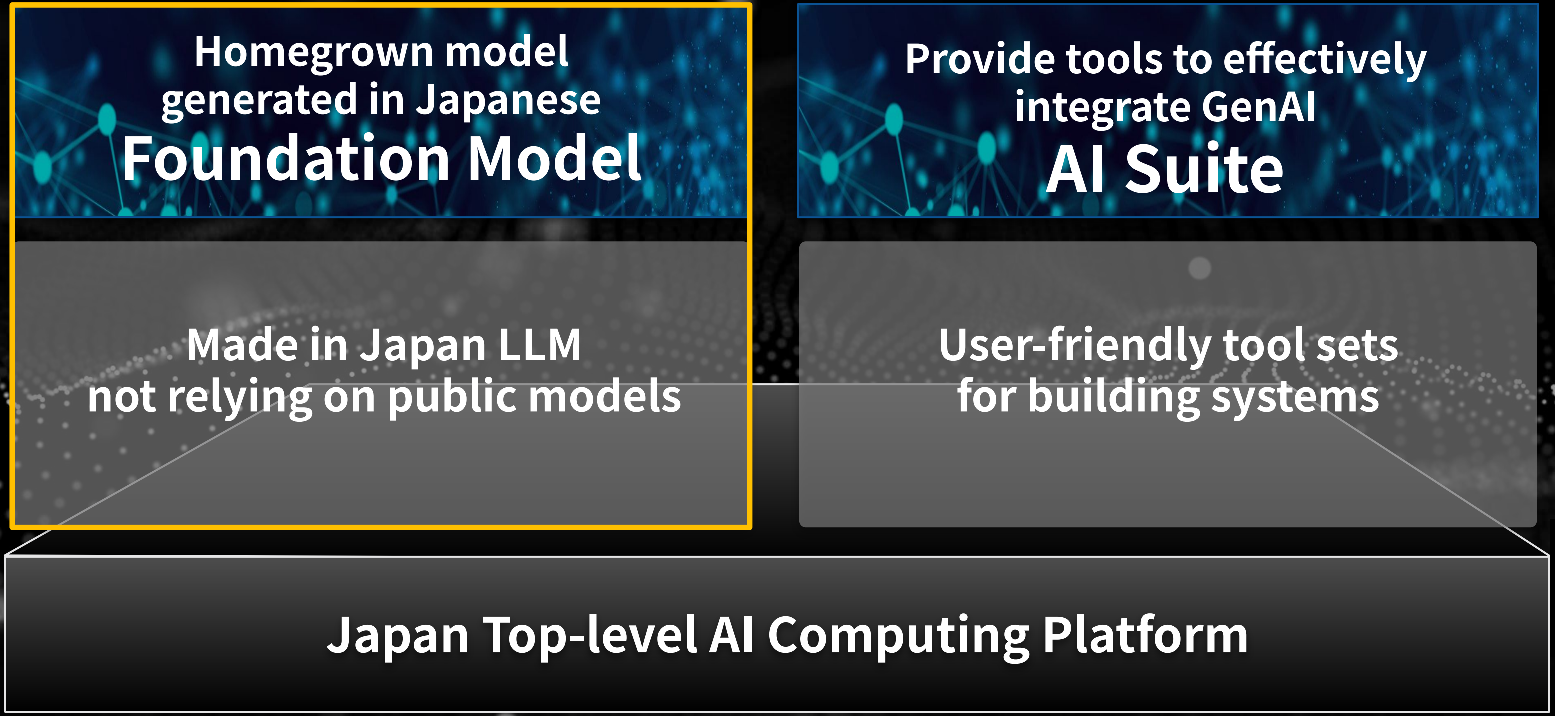
FY 2023 Start Infrastructure

FY 2024 approx. 390B parameters

FY 2025〜 Further Strengthening 1T parameters

[Notes] Largest computing platforms in Japan for LLMs learning. Based on publicly available information as of October 31, 2023, and SoftBank Corp. research.

SB Intuitions' Foundation Model and AI Suite



Homegrown model generated in Japanese

Foundation Model

Made in Japan LLM not relying on public models

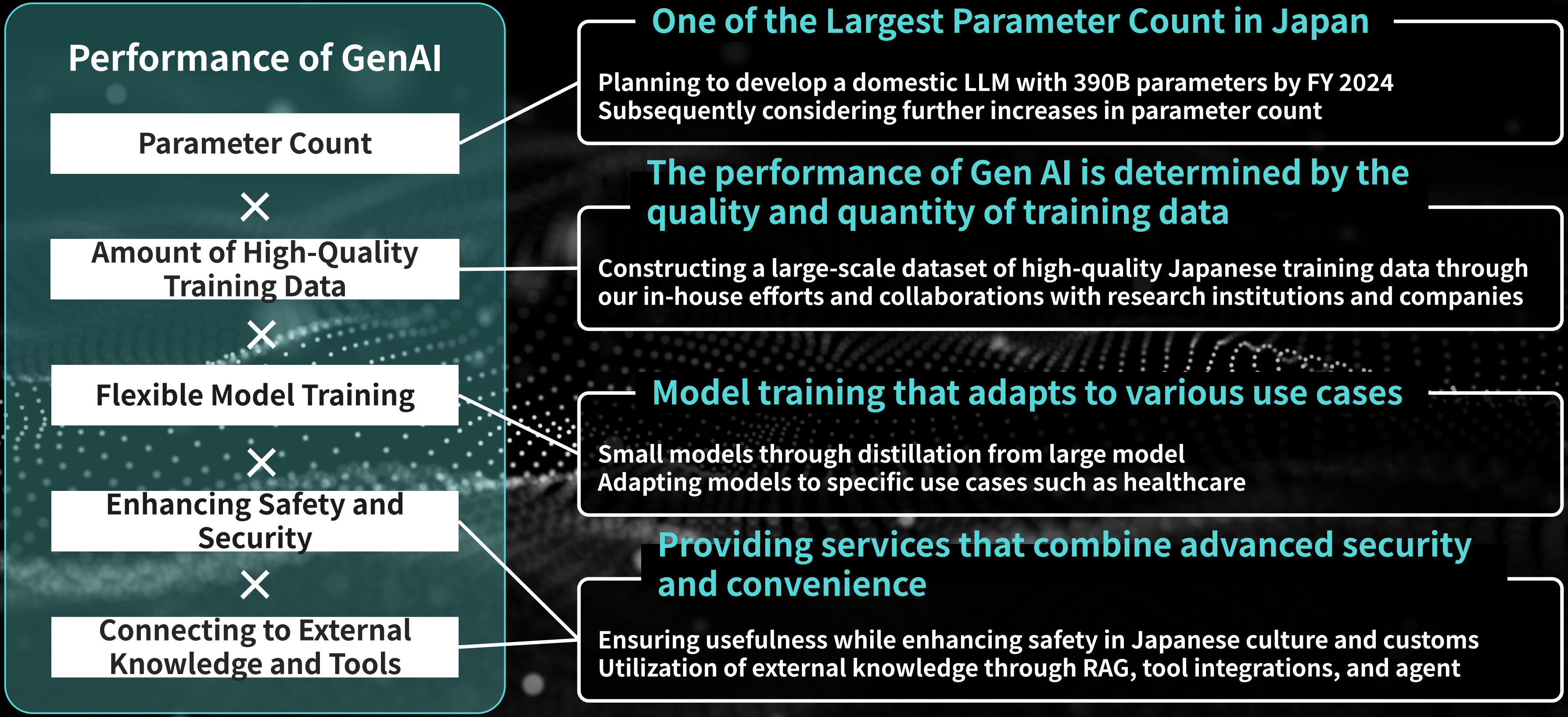
Provide tools to effectively integrate GenAI

AI Suite

User-friendly tool sets for building systems

Japan Top-level AI Computing Platform

Building an original Generative AI foundation model with high Japanese language performance



Performance of GenAI

Parameter Count

×

Amount of High-Quality Training Data

×

Flexible Model Training

×

Enhancing Safety and Security

×

Connecting to External Knowledge and Tools

One of the Largest Parameter Count in Japan

Planning to develop a domestic LLM with 390B parameters by FY 2024
Subsequently considering further increases in parameter count

The performance of Gen AI is determined by the quality and quantity of training data

Constructing a large-scale dataset of high-quality Japanese training data through our in-house efforts and collaborations with research institutions and companies

Model training that adapts to various use cases

Small models through distillation from large model
Adapting models to specific use cases such as healthcare

Providing services that combine advanced security and convenience

Ensuring usefulness while enhancing safety in Japanese culture and customs
Utilization of external knowledge through RAG, tool integrations, and agent

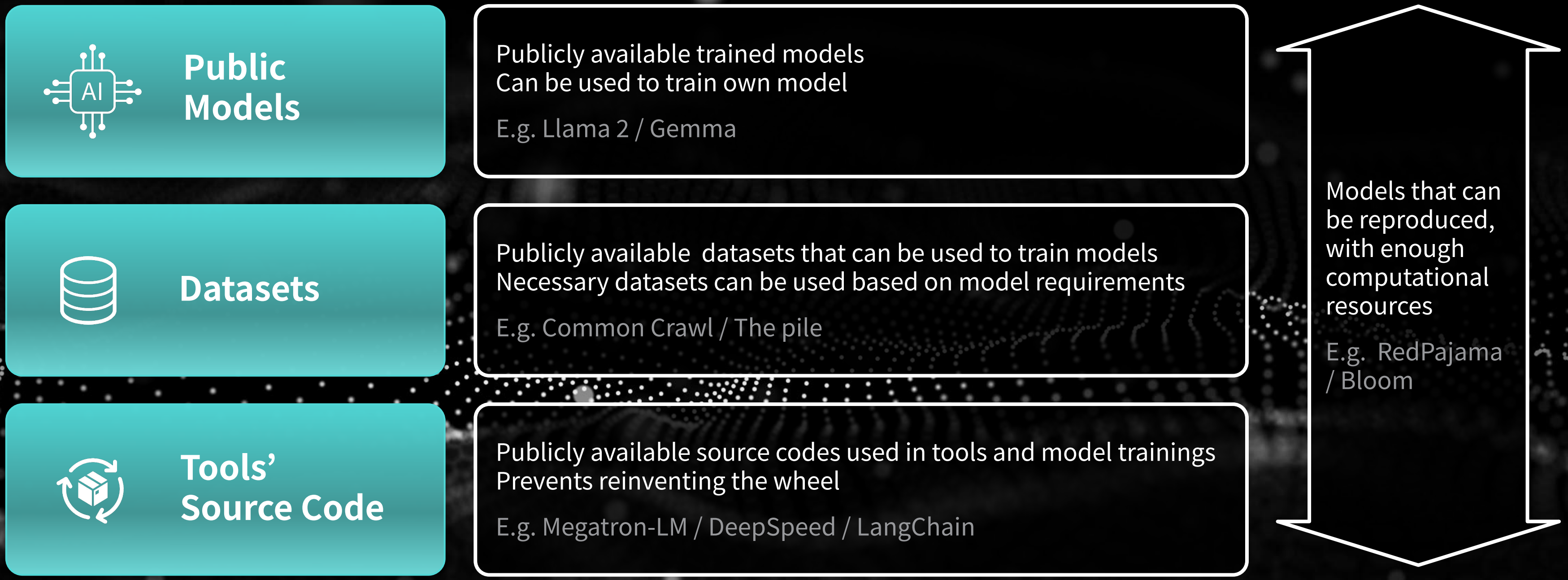
Develop a foundation model from scratch using Japanese data

Homegrown model with high Japanese language performance using large computational resources

	Continuous training on public models	Development not reliant on public models
Delivery	◎ **Low computational resources** Can quickly develop required models using existing public models	△ **Large computational and engineering resources** Extensive cost, time & human resources
Dataset	× **Not controllable** Issues in trusting datasets used by public models (copyrights etc.)	◎ **Flexibility to suit needs** Controllable datasets that can be adapted to the needs
Expansion	△ **Dependent on model provider** Safety issues in using models that are not continuously thoroughly checked	◎ **Stability can be preserved** Own model and datasets promote continuous R&D for expansion
Model quality	○ **Already high performing models** Use existing high performing general models as base	◎ **Specialization to suit needs** Filtering unnecessary data to increase the quality of specialized models

Breakdown of GenAI-related open source data

Aside from public LLM Models, other open source repositories serving different purposes exist



Public Models

Publicly available trained models
Can be used to train own model

E.g. Llama 2 / Gemma

Datasets

Publicly available datasets that can be used to train models
Necessary datasets can be used based on model requirements

E.g. Common Crawl / The pile

Tools' Source Code

Publicly available source codes used in tools and model trainings
Prevents reinventing the wheel

E.g. Megatron-LM / DeepSpeed / LangChain

Models that can be reproduced, with enough computational resources

E.g. RedPajama / Bloom

The competition in data acquisition

BigTech has a culture of litigations. Inequality in legally approved approach is widening.
A win-win relationship between content holders and AI developers is needed.

- Copyright law, article 30-4
- Agency for cultural affairs' "Thoughts on AI & Copyrights"
- G7 Hiroshima AI Process
- New business operators guidelines

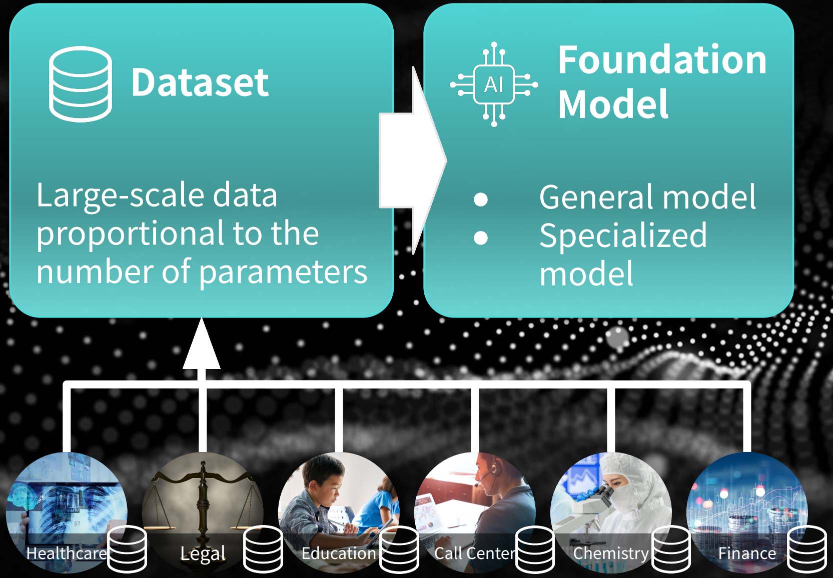
etc.

+

Rules and processes regarding data usage

GenAI is entering the competition to acquire data

Overwhelming shortage of data
Rapid need for a data circulation infrastructure



Dataset

Large-scale data proportional to the number of parameters

Foundation Model

- General model
- Specialized model

Hurdles to data circulation

Handling of personal information
Anonymization of personal information etc., process that satisfies end-users

Traceability
Can trace how data are used in learning and generation

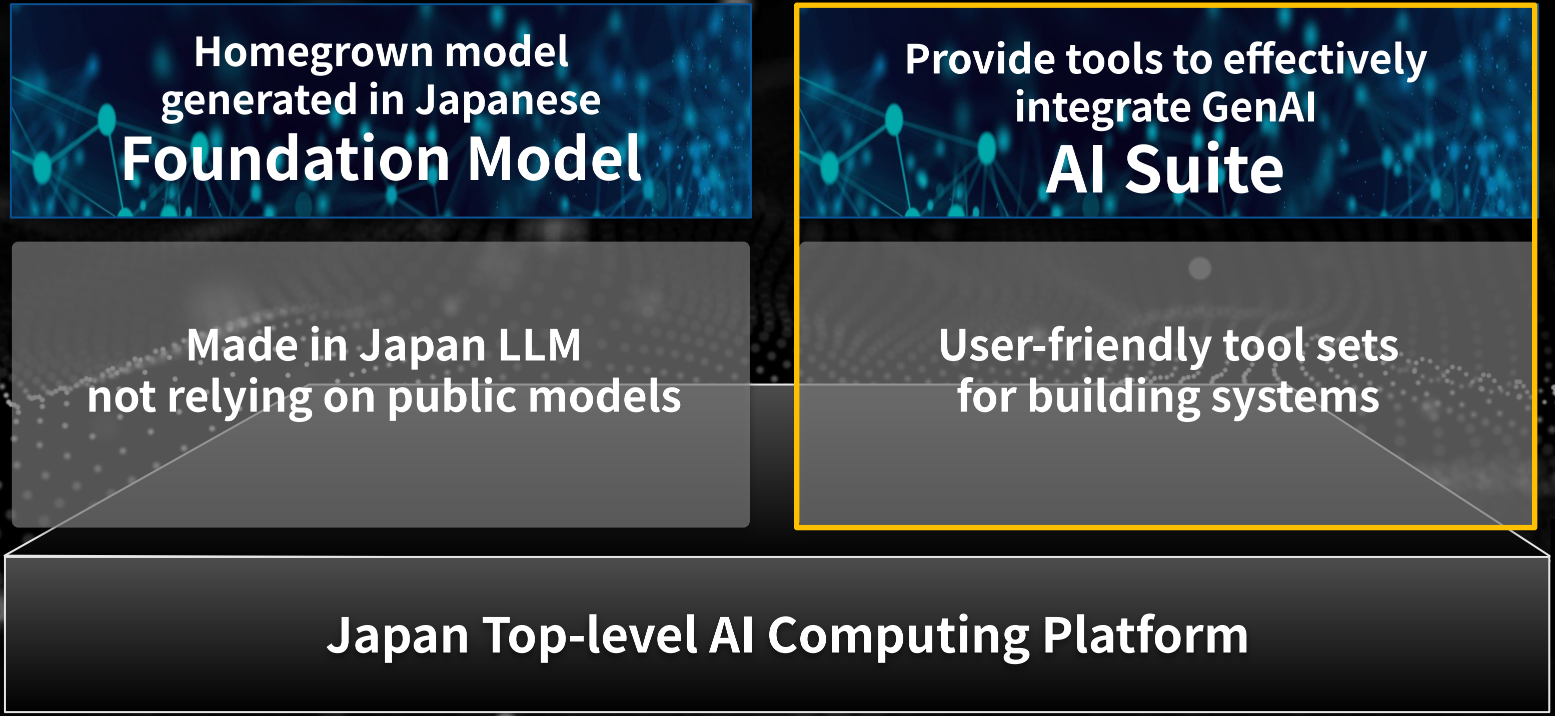
Preservation of copyrights etc.
Rules making for proper data handlings

Strategic alliances
Create general models using public data or differentiate itself

Returns to data holders
Return financial merits to data providers for long-term growth

Thorough usage controls
Process that ensures data circulation compliance

SB Intuitions' Foundation Model and AI Suite



Homegrown model generated in Japanese

Foundation Model

Made in Japan LLM not relying on public models

Provide tools to effectively integrate GenAI

AI Suite

User-friendly tool sets for building systems

Japan Top-level AI Computing Platform

AI Suite for customers to create their GenAI environment

We provide comprehensive support tools to ensure seamless integration of our foundation model into our customers’ systems

AI Suite

AI Solution Tools

Necessary toolkit for utilizing GenAI, including tools such as Agent and RAG

Model Store

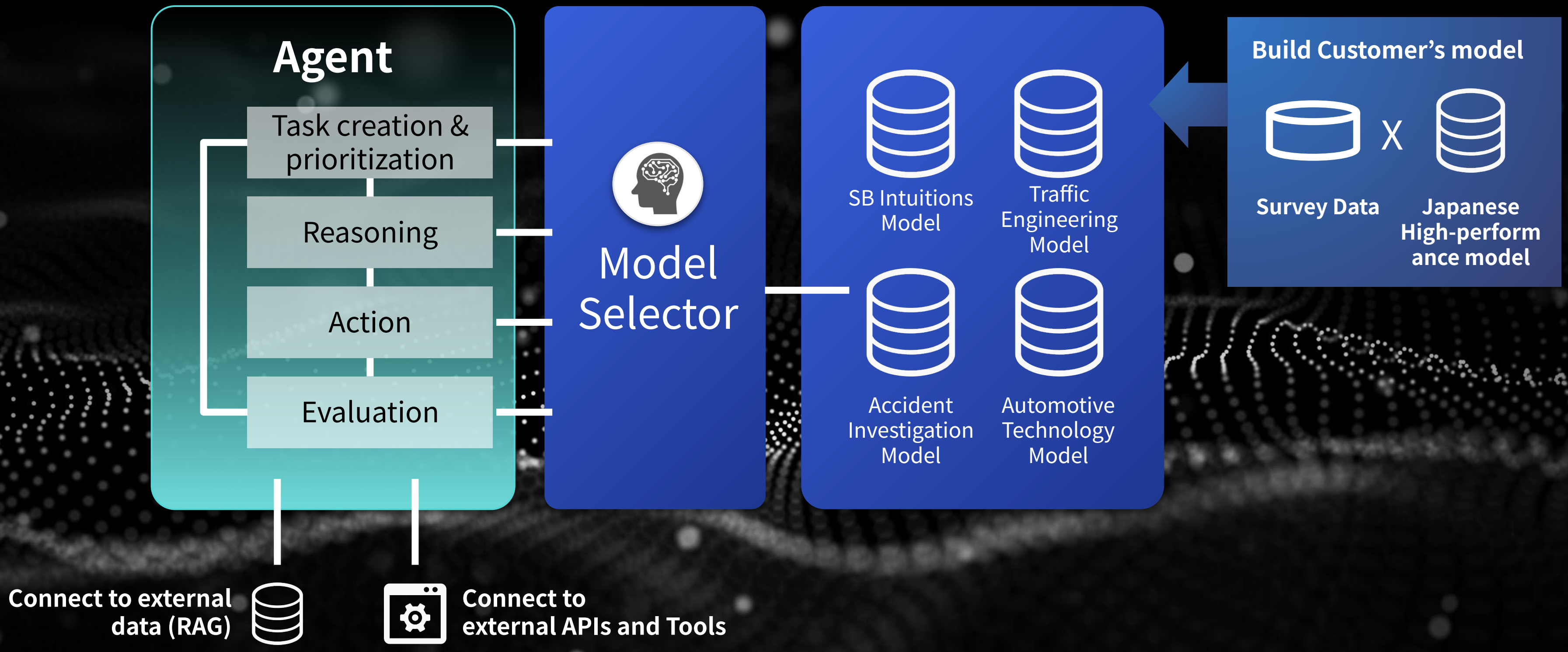
Provide a wide range of models, including high-performance models with excellent Japanese language capabilities and various expert models

ML Ops Tools

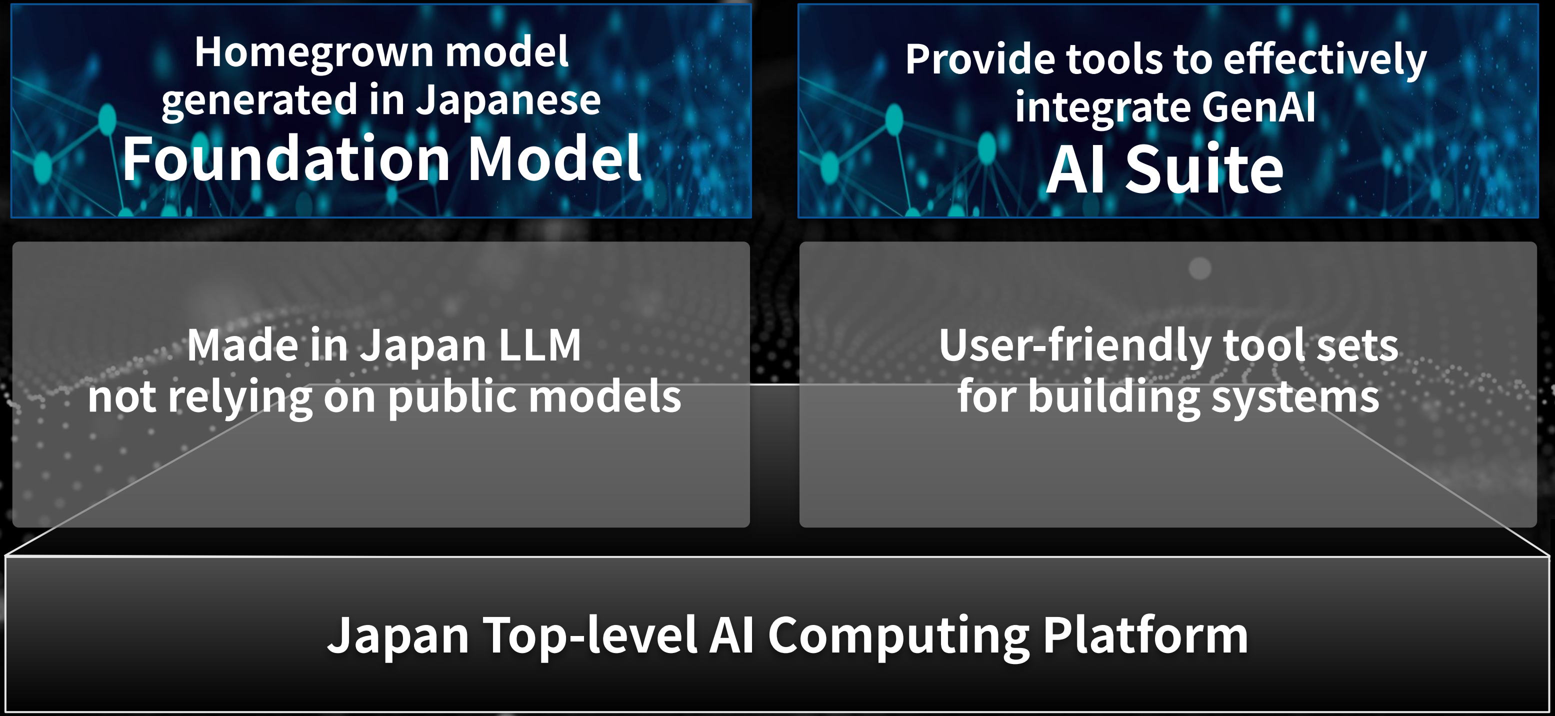
Seamless and comfortable UX for retraining and general operations under a robust data management system

Concept of incorporating multiple models into one system

Facilitate the opinions of multiple expert models and reaches conclusions autonomously


Agent
Task creation & prioritization
Reasoning
Action
Evaluation
Model Selector
SB Intuitions Model
Traffic Engineering Model
Accident Investigation Model
Automotive Technology Model
Build Customer's model
Survey Data
X
Japanese High-perform ance model
Connect to external data (RAG)
Connect to external APIs and Tools

SB Intuitions' Foundation Model and AI Suite



Homegrown model generated in Japanese

Foundation Model

Provide tools to effectively integrate GenAI

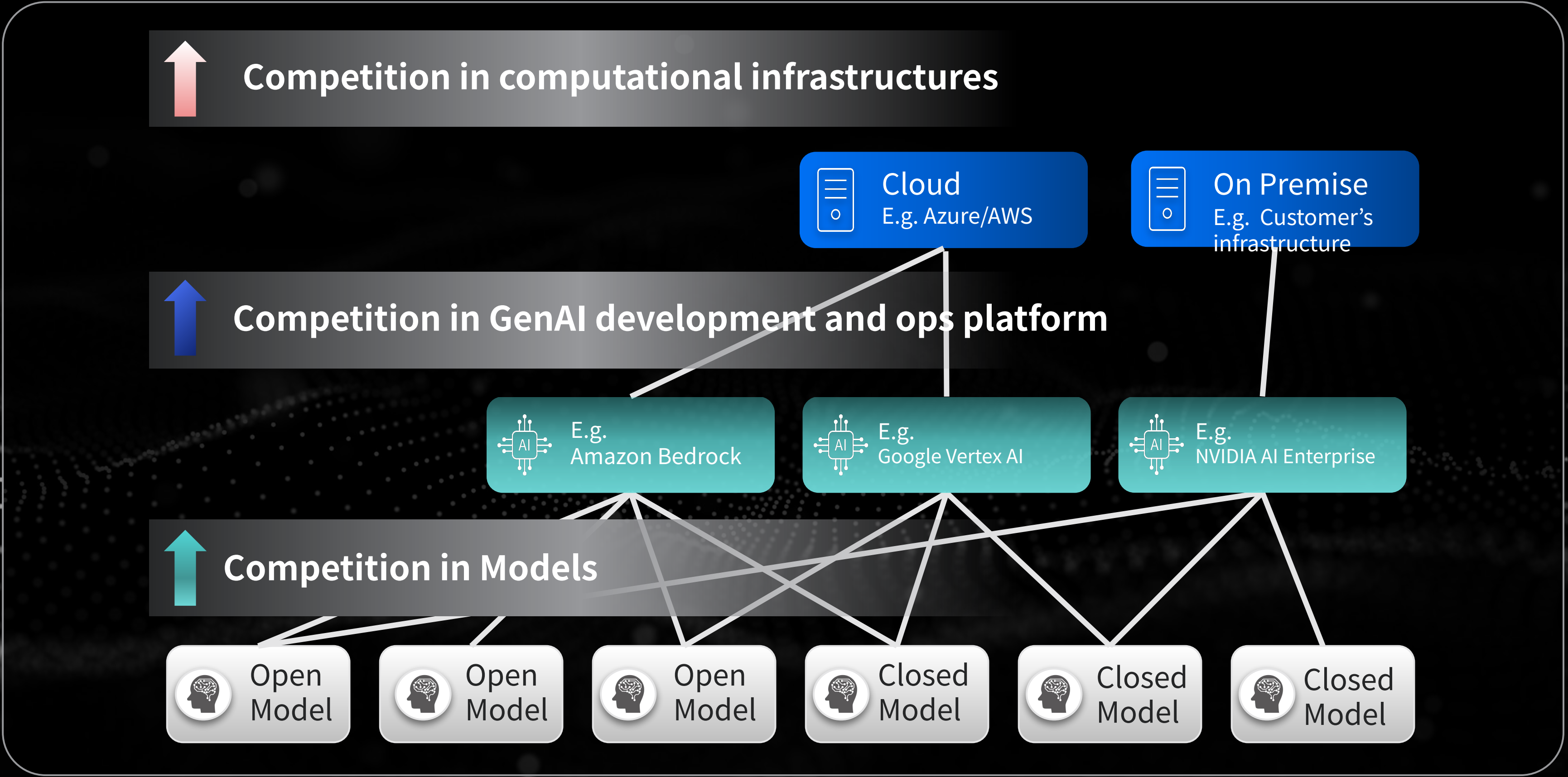
AI Suite

Made in Japan LLM not relying on public models

User-friendly tool sets for building systems

Japan Top-level AI Computing Platform

Competitive environment surrounding GenAI



Competition in computational infrastructures

Cloud
E.g. Azure/AWS

On Premise
E.g. Customer's infrastructure

Competition in GenAI development and ops platform

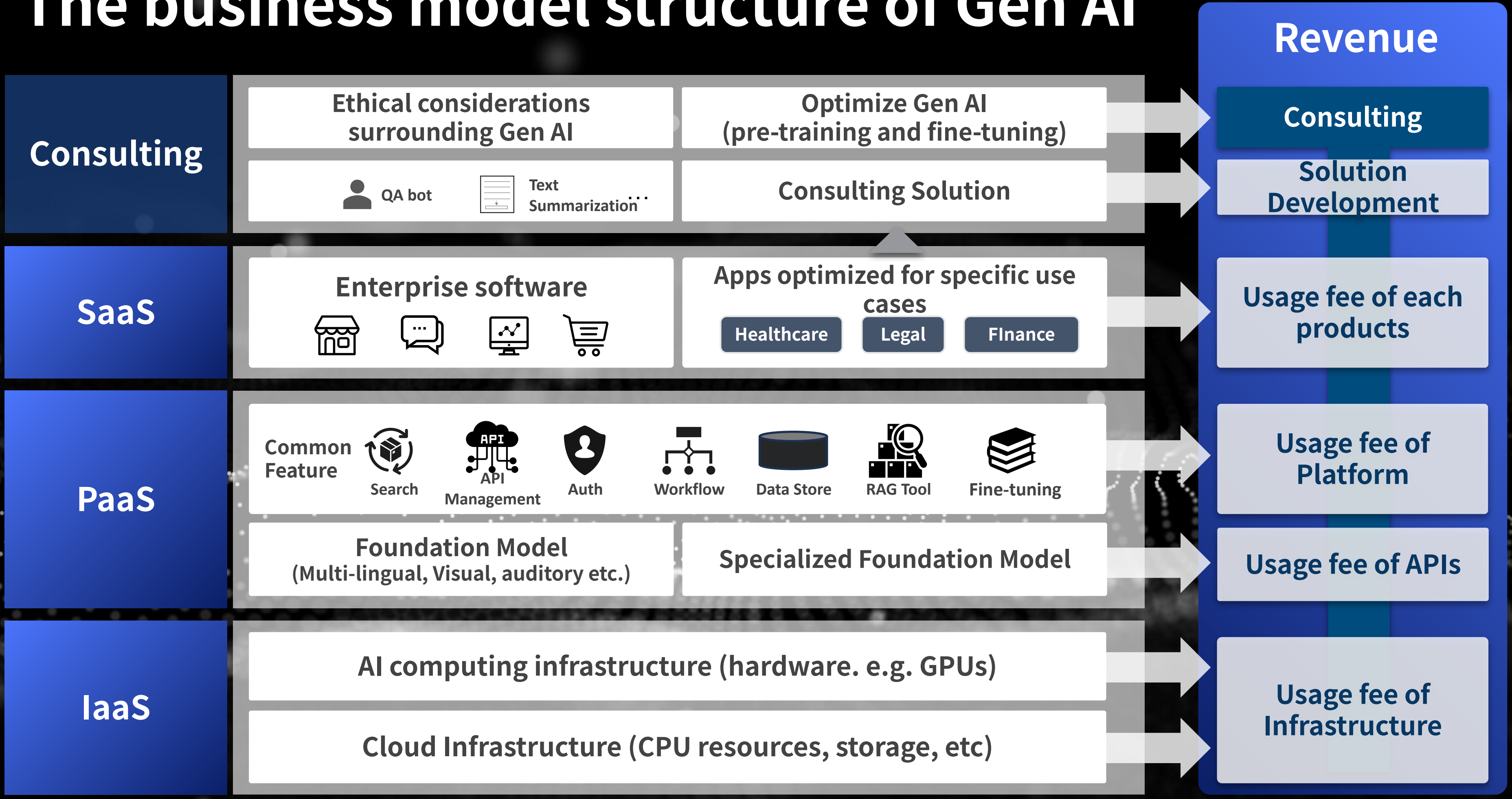
E.g.
Amazon Bedrock

E.g.
Google Vertex AI

E.g.
NVIDIA AI Enterprise

Competition in Models

The business model structure of Gen AI



Consulting

Ethical considerations surrounding Gen AI

Optimize Gen AI (pre-training and fine-tuning)

QA bot

Text Summarization ...

Consulting Solution

SaaS

Enterprise software

Apps optimized for specific use cases

Healthcare

Legal

FInance

PaaS

Common Feature

Search

API Management

Auth

Workflow

Data Store

RAG Tool

Fine-tuning

Foundation Model (Multi-lingual, Visual, auditory etc.)

Specialized Foundation Model

IaaS

AI computing infrastructure (hardware. e.g. GPUs)

Cloud Infrastructure (CPU resources, storage, etc)

Revenue

Consulting

Solution Development

Usage fee of each products

Usage fee of Platform

Usage fee of APIs

Usage fee of Infrastructure

SB Intuitions’ Vision

SB Intuitions' GenAI serves as a competitive advantage across all industries, fueling growth and progress in society

Healthcare

Legal

Education

Call Center

Chemistry

Finance

Foundation Model

AI Suite

Japan Top-level AI Computing Platform

SB Intuitions