AI Technology

The Potential of AI Under Labor Supply Constraints

May, 2025

Industry Research Department
Mizuho Bank

Summary

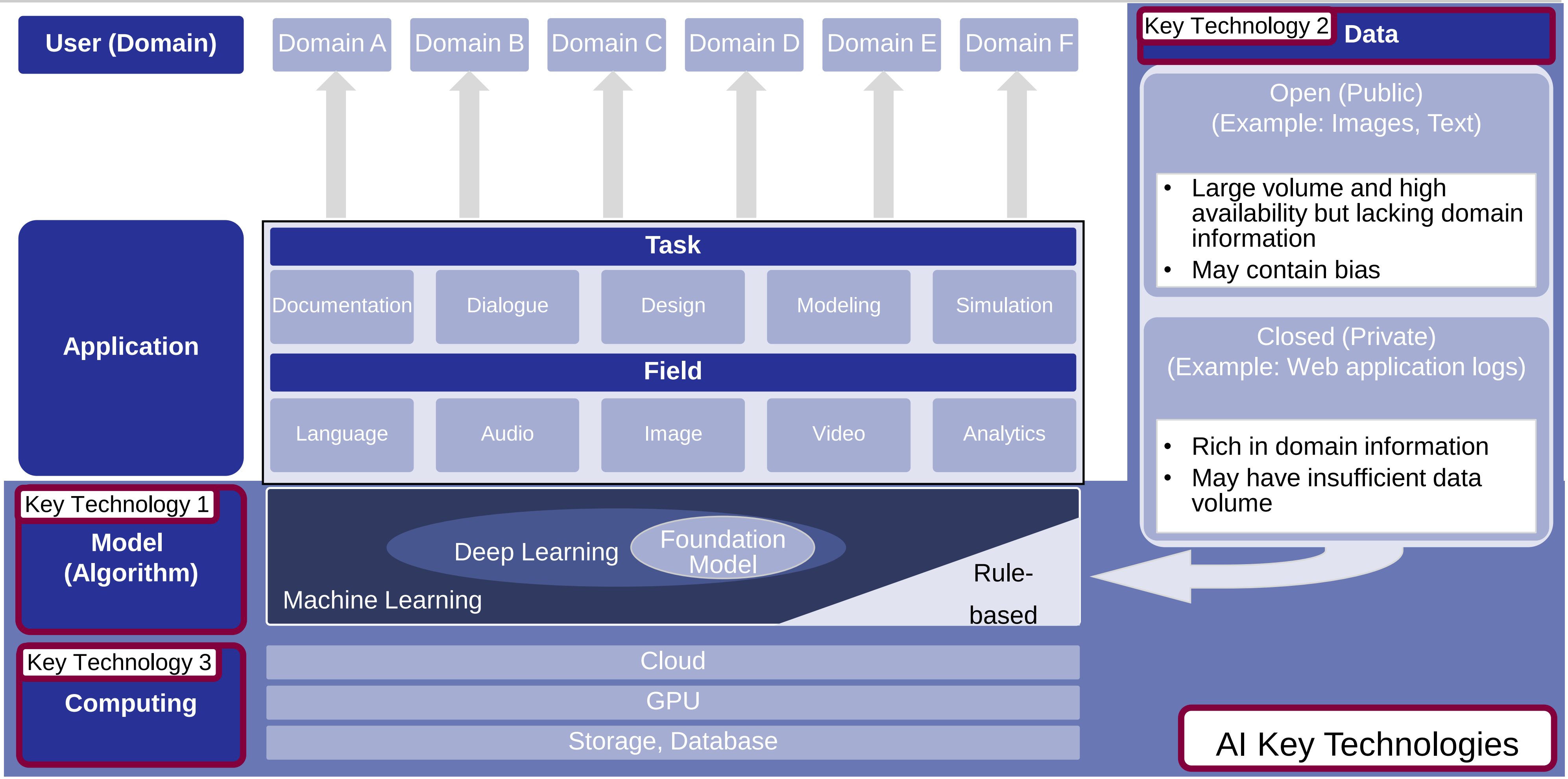
- AI consists of three key technologies: “models (algorithms),” “data,” and “computing,” and scaling laws, where performance improves as the scale of each key technology increases, have become the mainstream of AI development. This is based on the technological innovation of the Transformer, proposed by Google researchers in 2017, which has contributed to improving the generality and accuracy of AI models through scaling. However, AI faces challenges such as the “black box” problem and securing training data and computing resources.
- For the "black box" problem, combining classical AI (rule-based) systems, which have high explainability, with foundation models through AI hybridization is expected to provide solutions, potentially contributing to AI deployment in use cases requiring more advanced decision-making.
- When it comes to securing computing resources, there has been a shift from pursuing accuracy based on conventional scaling laws to exploring resource-efficient development methods (e.g., Mixture of Experts (MoE)), with the possibility of developing highly economically rational AI foundation models.
- With the resolution of technical challenges, AI is expected to contribute to addressing labor shortages. The key to utilizing AI for labor shortages is identifying AI deployment domains for each industry and developing and deploying appropriate AI applications.
- From a market perspective, the scale of the labor substitution market through AI applications is projected to reach approximately 34 trillion yen by 2050. However, in a scenario where deployment progresses gradually due to investment capacity constraints despite substitution being technologically possible, it would remain at approximately 16 trillion yen. The scale of the labor augmentation market is expected to peak around 2040 and reach approximately 562 billion yen by 2050, but under a gradual labor substitution scenario, it is expected to grow to approximately 949 billion yen.
- Scaling up closed data is key to developing and deploying appropriate AI applications. However, Japan's industrial structure and delayed DX may result in insufficient closed data for training data, creating barriers to AI deployment. While it would be desirable to centralize industry domain knowledge through data integration across companies to scale up closed data, actual data integration is expected to be difficult due to regulations, compliance, and competitive dynamics between companies.
- As a countermeasure to these barriers, federated learning may be technically effective. Federated learning is a technology for developing AI while data remains distributed across companies, enabling AI development when central data integration is difficult by scaling up closed data in a way that avoids data integration.

Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

AI Structure | Diverse applications and use cases are envisioned based on three key technologies

- The key technologies of machine learning are "models (algorithms)," "data," and "computing," and performance improves as each key technology becomes larger in scale.
  - — While applications are diverse, the areas that determine performance are the above three key technologies.

AI Architecture and Key Technologies



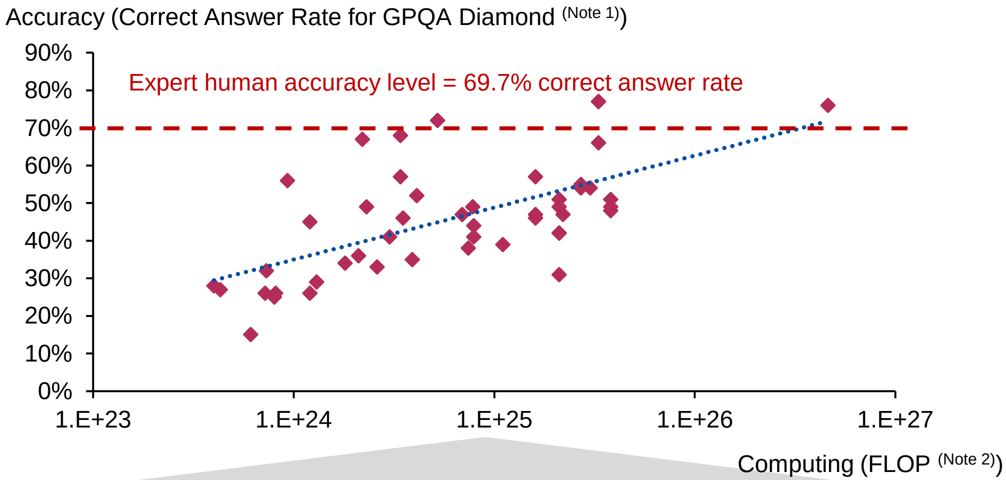
Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

Technological innovations in machine learning have improved accuracy and spread to business use

- Technological innovations in the layers of "models," "data," and "computing resources," which are the foundational technologies of machine learning, have enabled highly accurate responses in tasks such as image recognition and natural language processing.
- Generative AI is already being utilized for some white-collar work, with over 90% utilization in developed countries excluding Japan.
  - — Japan lags behind other developed countries in utilization, and while future utilization is expected, there may be challenges.

AI Performance Evolution and Innovation in Foundational Technologies

Trends in AI Performance and Computing



Innovations that Created Large-scale Key Technologies

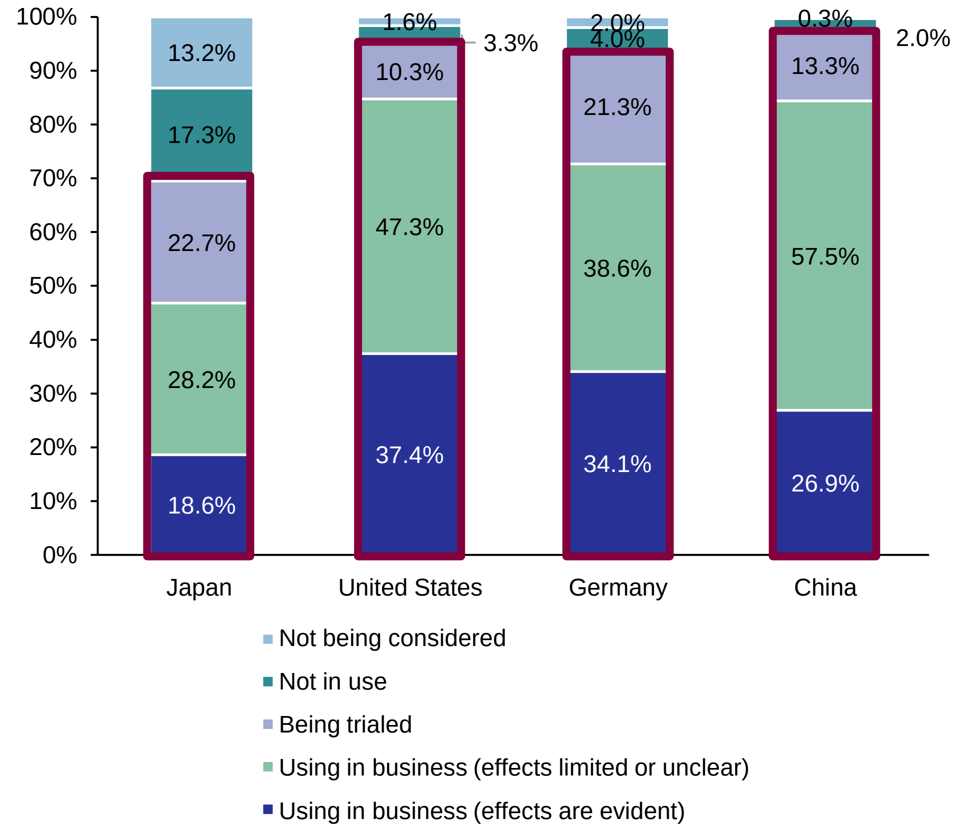
Model	Data	Computing
GPU parallel processing through Transformer/Google (2017)	Training data development through self-supervised learning	Computing speed improvement through GPU performance enhancement

Note 1 Benchmark for evaluating AI systems

Note 2 Abbreviation for "floating point operations," indicating the computational volume required for AI development (training)

Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on research by Epoch AI (Epoch AI, 'AI Benchmarking Hub'. Published online at epoch.ai. Retrieved from 'https://epoch.ai/data/ai-benchmarking-dashboard' [online resource]. Accessed 9 May 2025.)

Utilization Status of Generative AI in Business Operations (assistance with emails, meeting minutes, document creation, etc.)



Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on Ministry of Internal Affairs and Communications (2024) "Research Study on the Latest Information and Communication Technology Research and Development and Digital Utilization Trends in Japan and Overseas,"

MIZUHO

The Transformer has contributed to AI evolution, but deployment also faces challenges

- Since the emergence of the Transformer, the development of large-scale models has progressed, with significant improvements in "generality" and "accuracy."
- On the other hand, challenges include the “black box” nature that makes it difficult to explain the reasons for decisions, and the securing of training data and computing resources.

Technical Features, Contributions, and Challenges of the Transformer

Transformer

AI Model Using Attention as Core Technology

Deep Understanding of Data Structure Possible	Easy Parallel Processing with GPU Possible
✓ Mechanism that allows models to understand where to focus (Attention) within training data when processing, enabling interpretation of relationships between long-distance elements	✓ Previous models required sequential processing with poor training efficiency, but the Transformer enables parallel processing, accelerating training speed by using large amounts of GPU computing resources

Contributions

Improved Accuracy	Increased Scale of Key Technologies (Scaling)	Achieving Generality
✓ Improved accuracy in natural language processing, image recognition, analytics (e.g., protein structure prediction) through deep understanding of data structure	✓ Easy parallel processing enables large-scale machine learning for models that previously couldn't fit in memory, and/or couldn’t carry out calculations in a realistic timeframe. ✓ Discovery of scaling laws where the increased scale of key technologies leads to improved accuracy	✓ Acquiring general task knowledge and capabilities through large-scale machine learning in advance. ✓ Executing various tasks from language and image understanding/generation to complex problem-solving through reasoning by implementing additional machine learning for target tasks (downstream tasks)

Issues

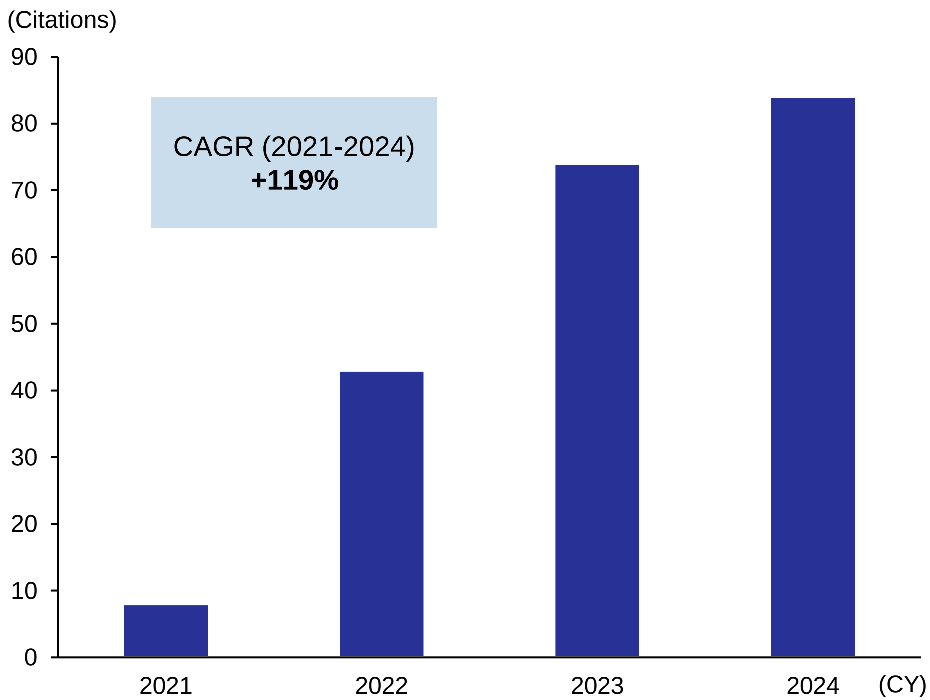
Trade-off between Accuracy and Explainability	Securing Training Data	Securing Computing Resources
✓ In machine learning, adopting deep neural networks with complex mathematical operations makes input-output causal relationships a “black box”, leading to difficulties in explaining AI decisions and their reasons	✓ Large-scale training data is necessary for improved AI accuracy through model scaling, and accuracy improvement reaches limits for use cases where scaling is difficult	✓ Large-scale AI development with models and training data requires massive computing resources. ✓ While improved accuracy from scaling models, data, and computing resources can be expected through "scaling laws," capital and revenue opportunities that are sufficient to cover development costs are necessary

Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

Progress in explainability research is expected to resolve the "black box" problem

- Research on concepts, technologies, and techniques to improve AI explainability is gaining attention.
  - For example, Neuro-symbolic AI, which combines machine learning/deep learning-based AI that excels in accuracy with rule-based models that excel in explainability, thereby resolving the “black box” problem.
- As the “black box” problem moves toward resolution, AI will become easier to use in use cases requiring more advanced decision-making, and AI's industrial deployment domain is expected to expand.

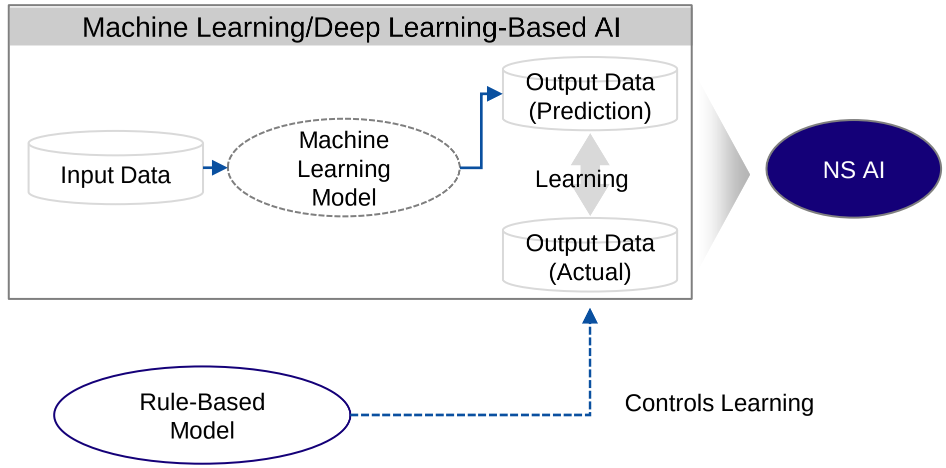
Number of Citations of Publications on Explainability Research (Example: Neuro-symbolic AI) (Note)



Note: Publications on explainability technology refer to Sarker, Md Kamruzzaman, et al. “Neuro-symbolic artificial intelligence: Current trends.” Ai Communications 34.3 (2022): 197-209. Citation count searched on Google Scholar (site accessed April 2025)

Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on Google Scholar

Resolving the “Black Box" Problem Through Neuro-symbolic AI



Neuro-symbolic AI (NS AI)

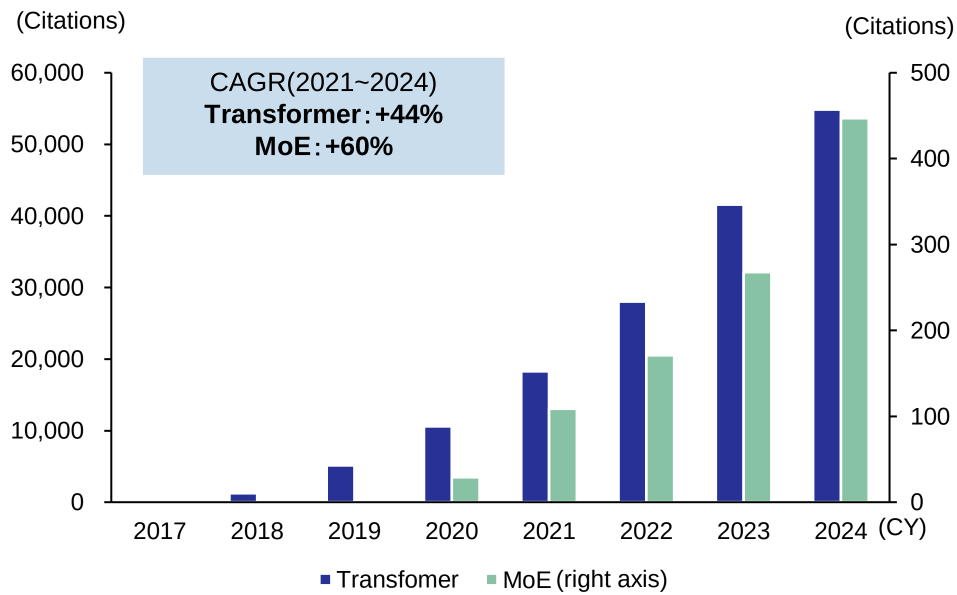
- ✓ When machine learning models learn from data, they learn through the control of information processing models (rule-based models) that are easy for humans to understand.
- ✓ Since rule-based models are reflected in learning, the “black box” is mitigated (improved explainability).

Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

The economic rationality of AI development is expected to improve through exploration of resource-efficient development methods

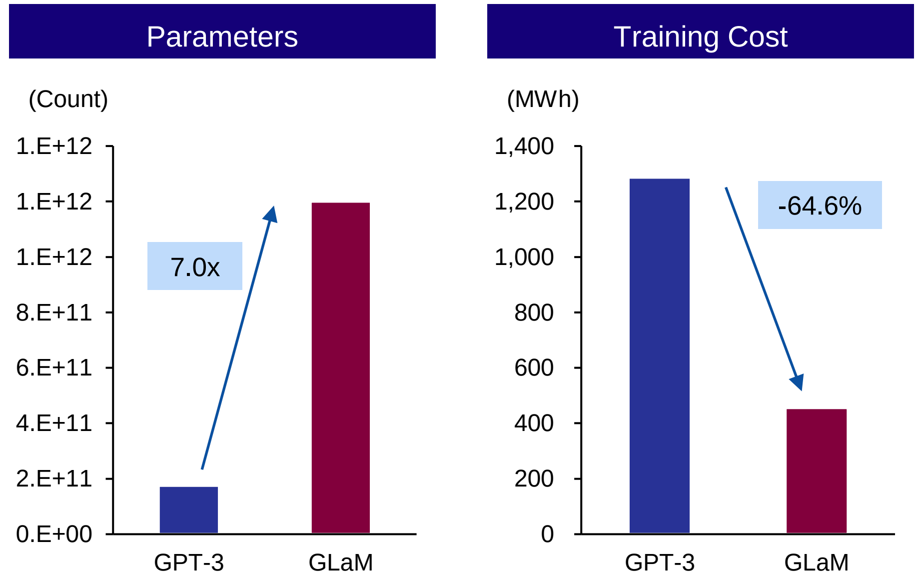
- Mainstream AI development is pursuing accuracy through large-scale models, data, and computing resources using Transformer technology.
  - However, resource-efficient development methods are also being explored (e.g., Mixture of Experts (MoE))
- Achieving large-scale models with relatively few computing resources.
  - The economic rationality of development is expected to improve in cases where computing resources are limited or it is difficult to expect commensurate business effects.

Number of Citations of Publications on the Transformer (left axis) and MoE (right axis) (Note 1, 2, 3)



Note 1: Publications on the Transformer refer to Vaswani, Ashish, et al. Attention is all you need. Citation count searched on Google Scholar (site accessed March 2025)
Note 2: Publications on MoE refer to Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., … & Chen, Z. (2020). Gshard: Scaling giant models with conditional computation and automatic sharding. Citation count searched on Google Scholar (site accessed March 2025)
Note 3: For MoE, refer to Mizuho Short Industry Focus Vol.246
Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on Google Scholar

Comparison between GLaM (Google) and GPT-3 (OpenAI)



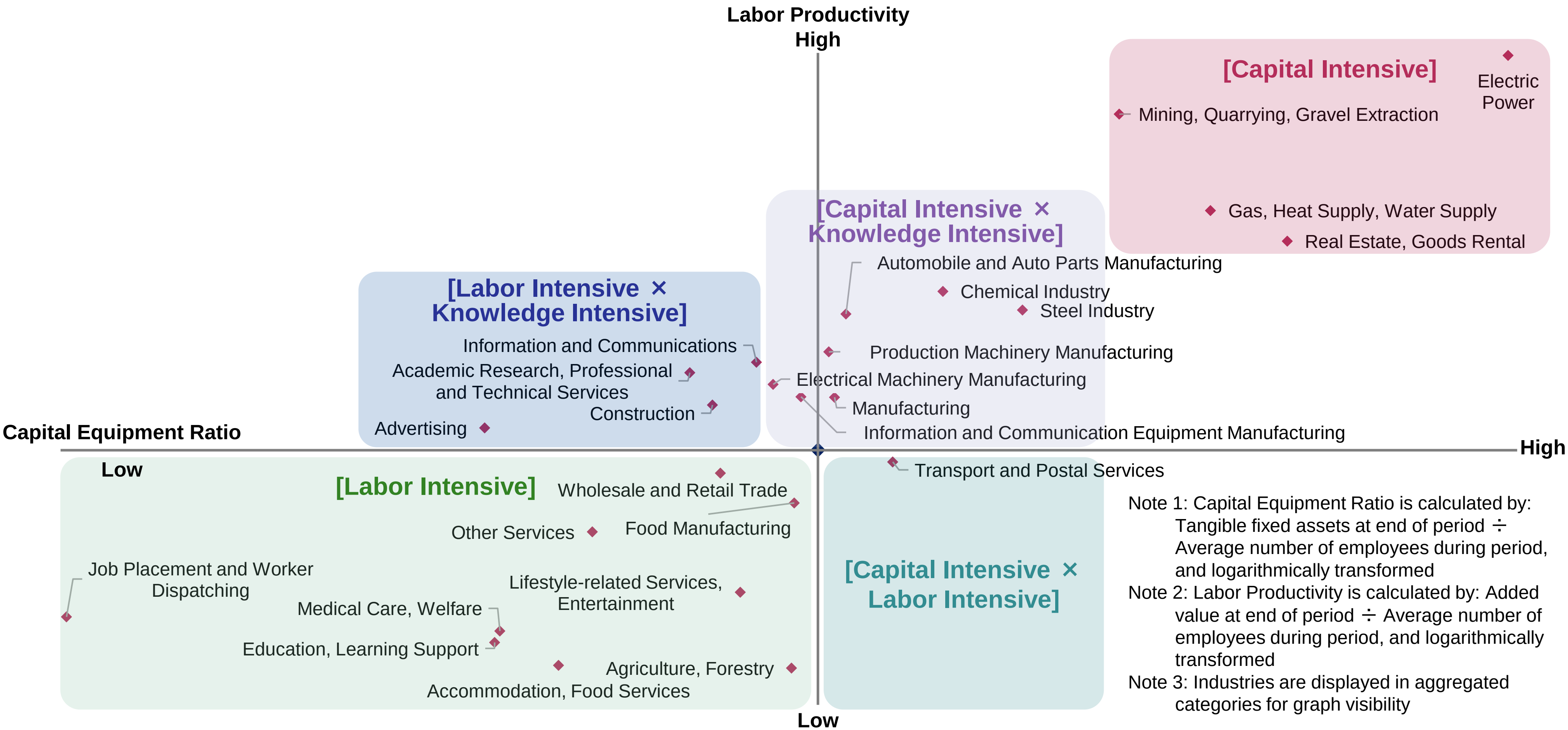
- Google released the AI model "GLaM" using MoE in 2021.
- GLaM has approximately 7 times the parameters of the then state-of-the-art AI model "GPT-3" (OpenAI), while being developed at lower cost than GPT-3.

Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on Du, Nan, et al. "Glam: Efficient scaling of language models with mixture-of-experts." International conference on machine learning. PMLR, 2022

AI deployment domains vary by motivation and can be classified according to industry types

- When aiming to resolve labor shortages through AI deployment, it is necessary to identify AI deployment domains for each industry and begin developing AI applications using closed data.
  - AI deployment domains are expected to differ according to industry type, based on industrial characteristics.

Industry Classification Based on Distribution of Capital Equipment Ratio and Labor Productivity by Industry



Note 1: Capital Equipment Ratio is calculated by: Tangible fixed assets at end of period ÷ Average number of employees during period, and logarithmically transformed
Note 2: Labor Productivity is calculated by: Added value at end of period ÷ Average number of employees during period, and logarithmically transformed
Note 3: Industries are displayed in aggregated categories for graph visibility

Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on Ministry of Finance "Financial Statements Statistics of Corporations by Industry (FY2023),"

Identify AI deployment domains and develop/deploy AI applications for problem-solving

- Since the areas with high labor cost burden rates differ by industry type, AI deployment domains are also expected to differ in preparing for the impact of labor shortages.
- First, it is necessary to identify areas requiring human labor and develop/deploy AI applications in those areas as a countermeasure against labor shortages.
  - To develop AI applications tailored to specific domains, it is necessary to secure closed data based on the problems to be solved.

Images of AI Deployment Domains and Specialized AI Applications by Industry Type

Labor Intensive × Knowledge Intensive (Information & Communications, Advertising)

Images of Initial AI Deployment Domains and Specialized AI Applications: Planning & Development → Production → Sales

- ➢ Content based on advanced expertise is the source of added value, with high labor cost burden rates in the planning stage
- ⇒ **Streamlining planning and development operations using user data (reducing workload)**

Capital Intensive × Knowledge Intensive (Manufacturing, Chemical)

Images of Initial AI Deployment Domains and Specialized AI Applications: R&D → Manufacturing → Sales

- ➢ Advanced products created through R&D are the source of added value, with high labor cost burden rates in R&D and manufacturing processes
- ⇒ **Advancing R&D through higher-value material exploration**
- ⇒ **Automating manufacturing operations using manufacturing process workflow data (reducing workload)**

Capital Intensive (Electric Power, Real Estate)

Images of Initial AI Deployment Domains and Specialized AI Applications: Procurement → Manufacturing → O&M

- ➢ Infrastructure requiring substantial capital investment is the source of added value, with high labor cost burden rates in infrastructure O&M
- ⇒ **Optimizing maintenance using data generated from infrastructure (reducing workload)**
- ⇒ **Converting tacit knowledge into organizational knowledge (advancing technical succession)**

Labor Intensive (Service, Retail)

Images of Initial AI Deployment Domains and Specialized AI Applications: Procurement → Distribution → Service Sales

- ➢ On-site service provision and product sales are the source of added value, with high downstream labor cost burden rates
- ⇒ **Streamlining administrative work using workflow data from non-face-to-face operations (reducing workload)**
- ⇒ **Business model transformation based on digital premises**

Capital Intensive × Labor Intensive (Transportation)

Images of Initial AI Deployment Domains and Specialized AI Applications: Procurement → Service → Maintenance

- ➢ Asset-based service provision is the source of added value, with high labor cost burden rates in services and maintenance
- ⇒ **Automating operations using workflow data (labor-saving in services)**
- ⇒ **Optimizing maintenance using data generated from assets (reducing workload)**

Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

Market Size Concept: The amount companies are willing to pay for AI applications depends on their deployment motivation (substitution/augmentation)

- For labor shortages that are technologically substitutable/non-substitutable, AI applications (including those using robots) are deployed with business substitution/augmentation as motivation.
  - — The amount an organization is willing to pay for substitution/augmentation is equivalent to personnel costs/software expenses.

AI Market Size Concept

1 Market Definition	2 Scenario Assumptions
✓ The market is assumed to consist of two components: a "substitution market" based on demand motivated by job substitution and an "augmentation market" based on demand motivated by job augmentation.	✓ Two scenarios are assumed: one based on the feasibility of technological substitution ("Scenario A") and another based on the feasibility of substitution considering deployment lag(Note 1) ("Scenario B").

[Substitution Market]

- ➢ AI is deployed for substitutable labor(Note 2) during shortages, motivated by job substitution.
- ➢ An equivalent to personnel costs(Note 3) is paid for deployment.
  - — Personnel costs that would have been incurred without shortages are allocated to AI deployment.

[Augmentation Market]

- ➢ AI is deployed for non-substitutable labor during shortages, motivated by job augmentation.
- ➢ An equivalent to software expenses(Note 4) (per worker) is paid for deployment.
  - — Software expenses that would have been incurred without shortages are allocated to AI deployment.

Number of substitutable workers during 20xx labor shortage (Scenario A | Scenario B)
✖
Personnel costs
→ 20xx AI-driven labor substitution market (Scenario A | Scenario B)

Number of non-substitutable workers during 20xx labor shortage (Scenario A | Scenario B)
✖
Software expenses (per worker)
→ 20xx AI-driven labor augmentation market (Scenario A | Scenario B)

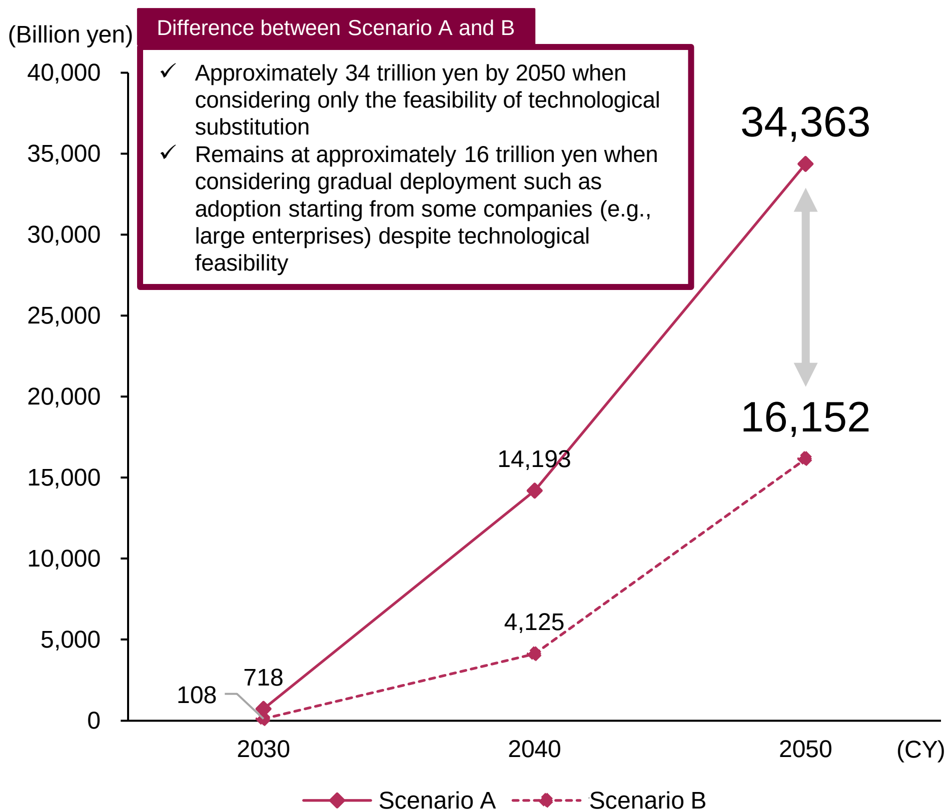
→ 20xx AI market size (Scenario A | Scenario B)

Note 1: For scenarios based on the feasibility of substitution considering deployment lag, refer to the assumptions of robot/AI deployment progress rates in the comprehensive edition
Note 2: For substitutable (non-substitutable) labor, refer to the comprehensive edition
Note 3: Personnel costs are calculated by dividing employee salaries (end of period) by the average number of employees during the period (end of period) for all industries (excluding finance and insurance), based on the Financial Statements Statistics of Corporations by Industry (FY2023)
Note 4: Software expenses (per worker) are calculated by dividing software (end of period fixed assets) by the average number of employees during the period (end of period) for all industries (excluding finance and insurance), based on Financial Statements Statistics of Corporations by Industry (FY2023), further divided by 5 years as useful life
Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

AI labor substitution market size will reach approximately 34 trillion yen by 2050

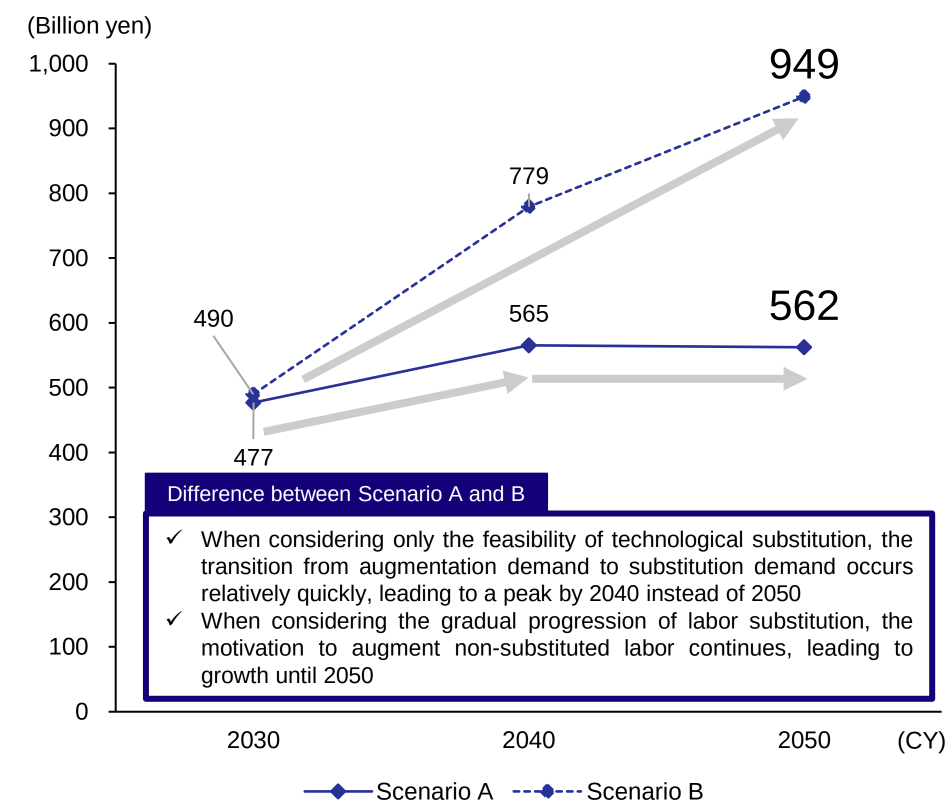
- The AI application market consists of two components: deployment motivated by "labor substitution" and deployment motivated by "labor augmentation."
  - The market size for labor substitution is projected to reach approximately 34 trillion yen by 2050 when considering only the feasibility of technological substitution ("Scenario A"). However, when considering deployment lag despite technological feasibility ("Scenario B"), it remains at approximately 16 trillion yen by 2050.
  - The market size for labor augmentation is expected to peak by 2040 and to reach approximately 562 billion yen by 2050 in Scenario A, as it transitions relatively quickly from augmentation demand to substitution demand. On the other hand, in Scenario B, it continues to increase until 2050, expanding to approximately 949 billion yen.

AI Market Size: Labor Substitution



Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

AI Market Size: Labor Augmentation

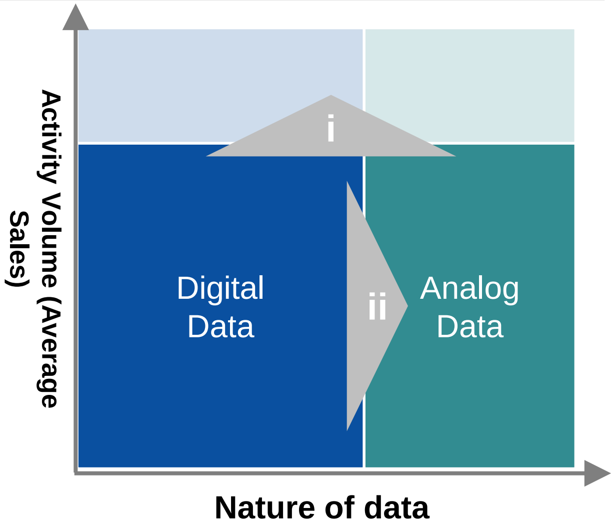


Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

Shortage of digital data due to industrial structure and delayed ICT investment may pose barriers

- In Japan, there are concerns about a shortage of closed digital data due to the industrial structure and delays in DX, which could pose barriers to building AI that has learned domain knowledge.
  - i. The average sales of domestic listed companies are one-third that of the US and about half that of China and the EU, with a relatively small scale per company, potentially resulting in less total data generated per company.
  - ii. Compared with other developed countries, the growth rate of ICT investment is low, and DX for AI learning may not be progressing properly.

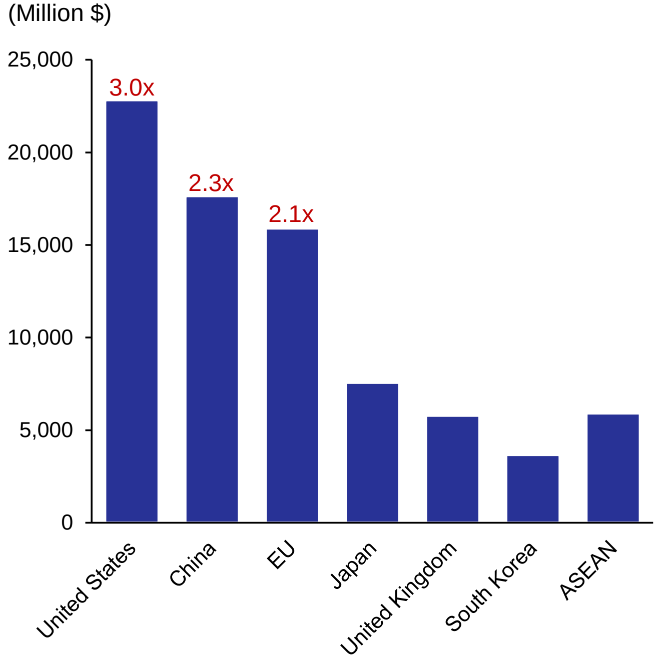
Concept of Closed Data for AI Deployment



➢ To scale up closed data, it is necessary to increase i. the amount of data per company(Note), ii. the proportion of digital data.

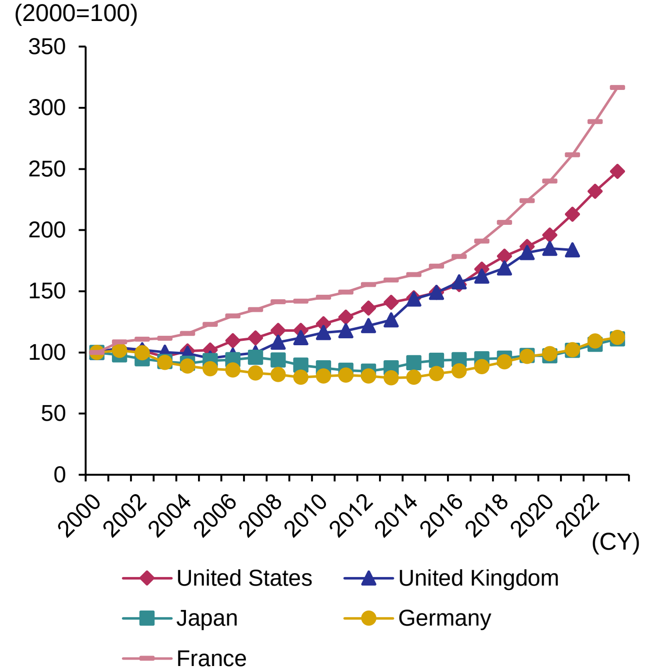
Note: The amount of data per company is assumed to be proportional to corporate activity volume (sales)
Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

i. Average Sales of Listed Companies by Country/Region



Note: Average sales of top 1,000 listed companies by sales in each country/region
Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on SPEEDA,

ii. ICT Investment Trends by Country



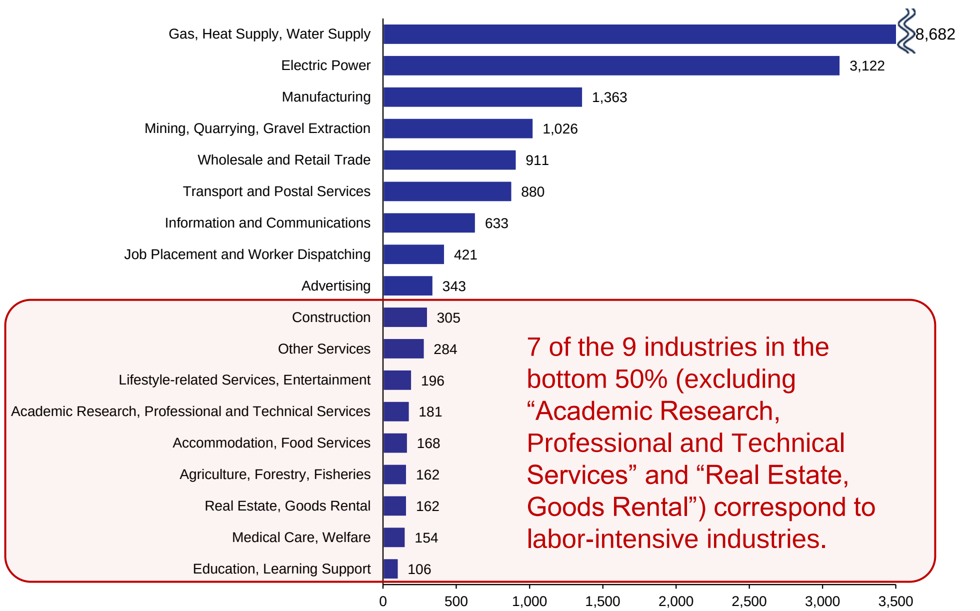
Note: Cumulative investment in ICT equipment, software, and databases
Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on OECD "Annual fixed assets by economic activity and by asset",

MIZUHO

Issues may be particularly pronounced in labor-intensive industries with significant impacts from labor shortages

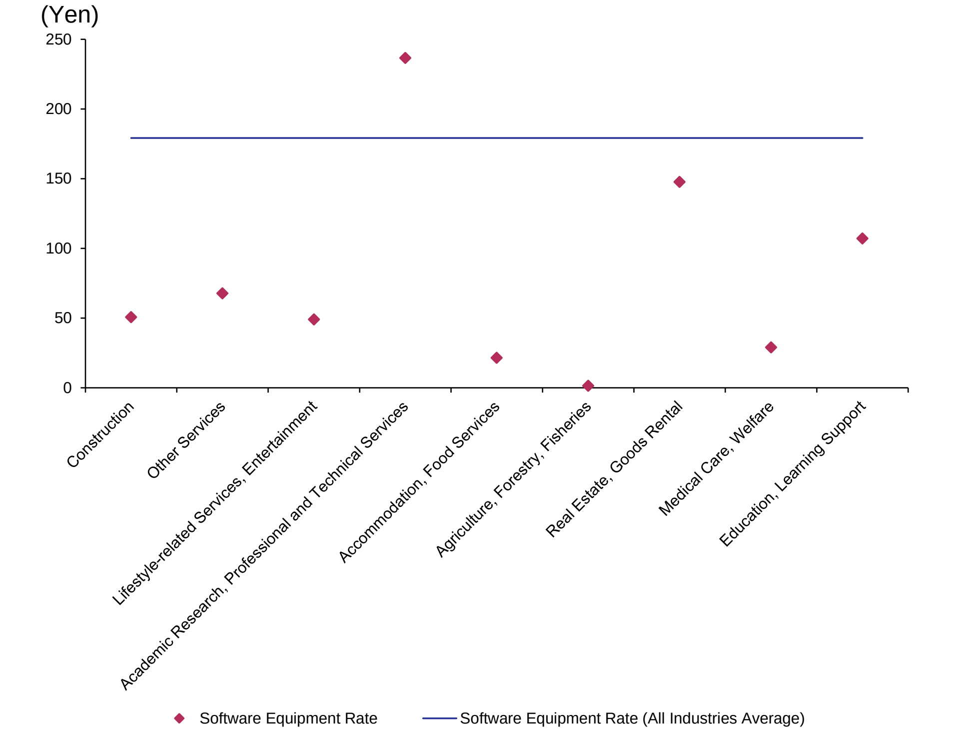
- The issue of digital data shortages due to small corporate scale and delayed DX is expected to be particularly prominent in labor-intensive industries in Japan.
  - — Labor-intensive industries in Japan are characterized by a smaller sales scale per company and lower software equipment rates compared with other industries.
- The labor shortages in labor-intensive industries in 2050 are expected to account for approximately 70% of Japan's total industrial labor shortage, necessitating consideration of countermeasures to resolve supply constraints.

Average Sales per Company by Industry in Japan (FY2023)



Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on Ministry of Finance “ Financial Statements Statistics of Corporations by Industry (FY2023),"

Software Equipment Rate for Industries in the Bottom 50% by Sales per Company



Note: Software Equipment Rate = Software Capital Stock ÷ Labor Input (Number of Employees × Working Hours)

Source: Compiled by Industry Research Department, Mizuho Bank, Ltd. based on Ministry of Finance " Financial Statements Statistics of Corporations by Industry (FY2023),"

MIZUHO

Federated learning becomes a countermeasure to overcome data scaling challenges

- When scaling up closed data, it is desirable to centralize industry domain knowledge through data integration across companies, but actual data integration is difficult due to regulations, compliance, and competitive dynamics between companies.
- For AI to contribute to resolving labor shortages, it is necessary to overcome the challenges of scaling up closed data, and federated learning is effective for this.
  - — This technology enables the construction of AI models while data remains distributed across organizations/companies, potentially serving as a countermeasure when data collaboration is difficult.

Countermeasures and Challenges for Scaling Up Closed Data

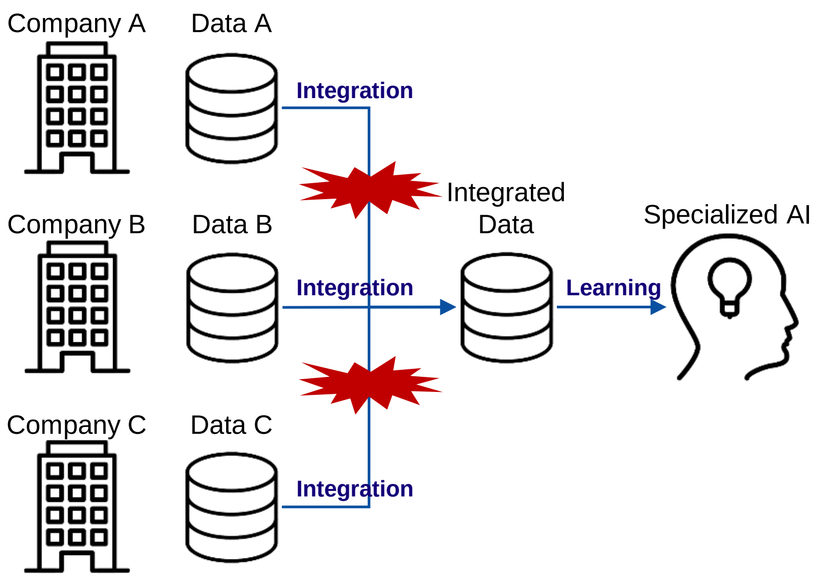
Ideal

- ➢ Integrate data across companies and scale up closed data to train industry-specialized AI

Gap

Reality

- ➢ Data integration is difficult due to regulations, compliance, competition with competitors, etc.



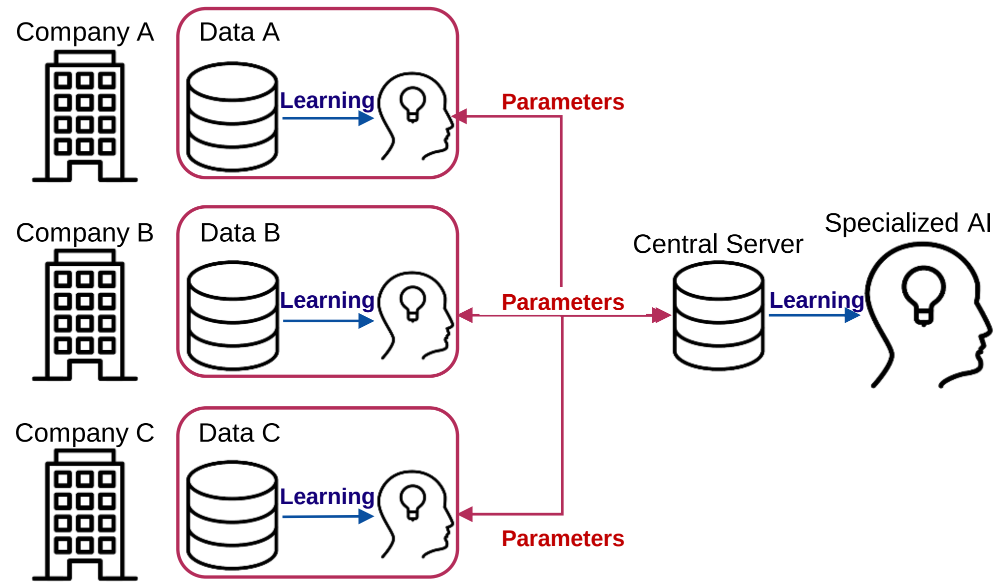
Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

Building Industry-Specialized AI Models Using Federated Learning Technology

Technical Features of Federated Learning

- ➢ Train local AI within each company's server and collect only learning results (parameters) in the central server to train industry-specialized AI.

⇒ It is possible to improve model accuracy while keeping data distributed



Source: Compiled by Industry Research Department, Mizuho Bank, Ltd.

Industry Research Department
Next-Generation Business Support Office
Strategic Project Team

Yuki Saito yuki.c.saito@mizuho-bk.co.jp

Yu Maeshima yu.maeshima@mizuho-bk.co.jp

Mizuho Industry Research／78 2025

Published May 30, 2025

Edited / issued by Industry Research Department, Mizuho Bank, Ltd 1-3-3 Marunouchi, Chiyoda-ku, Tokyo ird.info@mizuho-bk.co.jp