Press release

September 8, 2025
Fujitsu Limited

Fujitsu develops generative AI reconstruction technology for optimized and energy-efficient AI models based on Takane LLM

Achieving world's highest accuracy retention rate of 89% and 3x faster inference speed with 1-bit quantization memory consumption reduction of 94%

Kawasaki, Japan, September 8, 2025 – Fujitsu today announced the development of a new reconstruction technology for generative AI. The new technology, positioned as a core component of the Fujitsu Kozuchi AI service, will strengthen the Fujitsu Takane LLM by enabling the creation of lightweight, power-efficient AI models.

Fujitsu's new reconstruction technology is built upon two core advancements:

1. Quantization: A technique that significantly compresses information stored in the connections between neurons that forms the basis of an AI model's "thought" process
2. Specialized AI distillation: A world-first (1) method which simultaneously achieves both lightweighting and accuracy exceeding that of the original AI model

Applying 1-bit quantization technology to Takane has enabled a 94% reduction in memory consumption. This advancement has achieved the world's highest accuracy retention rate of 89% (2) compared to the unquantized model, along with a 3x increase in inference speed. This significantly surpasses the accuracy retention rate of less than 20% typically achieved by conventional mainstream quantization methods like GPTQ. This breakthrough enables large generative AI models that previously required four high-end GPUs to run efficiently on a single low-end GPU.

This unprecedented lightweighting capability will enable the deployment of agentic AI on edge devices such as smartphones and factory machinery. This will lead to improved real-time responsiveness, enhanced data security, and a radical reduction in power consumption for AI operations, contributing to a sustainable AI society.

Fujitsu plans to sequentially offer customers globally trial environments for Takane with the applied quantization technology starting in the second half of fiscal year 2025. Furthermore, Fujitsu will progressively release models of Cohere's research open-weight Command A quantized using this technology, available via Hugging Face (3) starting today.

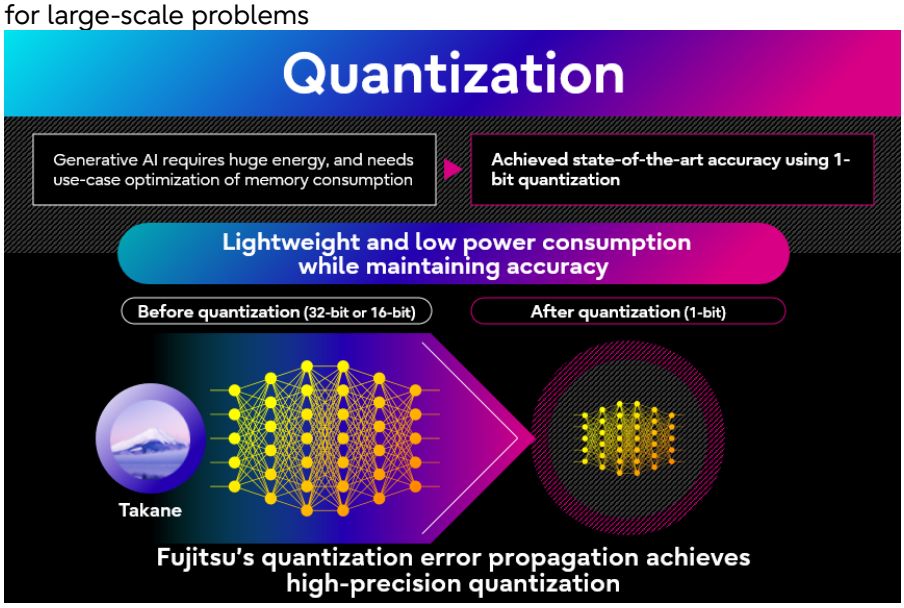
Moving forward, Fujitsu will continue to advance research and development that significantly improves the capabilities of generative AI while ensuring its reliability, aiming to solve more complex challenges faced by customers and society, and to open new possibilities for generative AI utilization.

Technology details

Many tasks performed by AI agents require only a fraction of the general capabilities of an LLM. The newly developed generative AI reconstruction technology is inspired by the human brain's ability to reconstruct itself, including by reorganizing neural circuits and specializing in specific skills in response to learning,

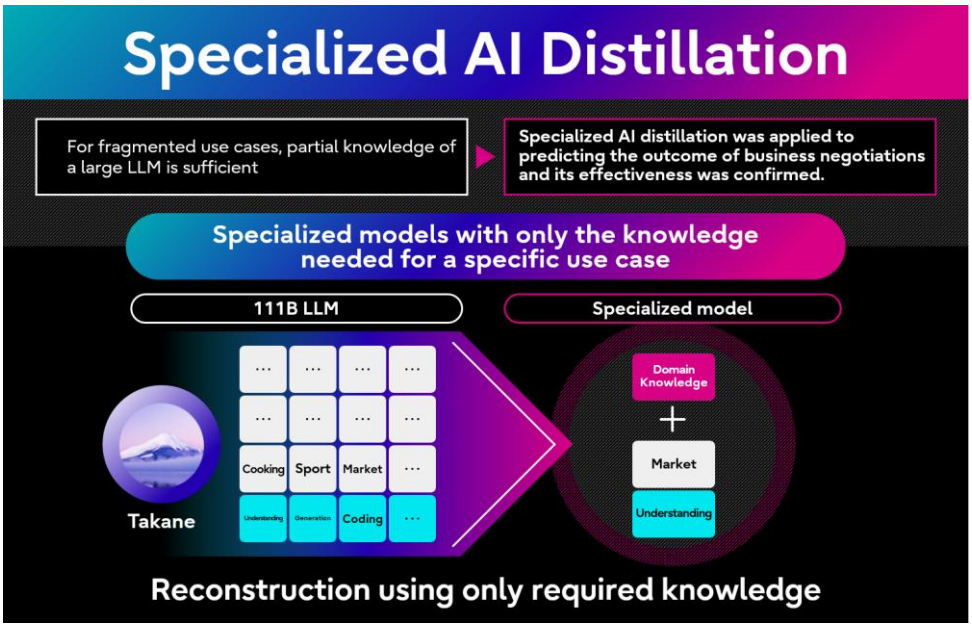
experience, and environmental changes. It efficiently extracts only the knowledge necessary for specific tasks from a massive model with general knowledge, creating a specialized AI model that is lightweight, highly efficient, and reliable. This is achieved through the following two core technologies:

1. **Quantization technology for streamlining AI "thought" and reducing power consumption:**
    - **Parameter compression:**
        - Technology compresses generative AI parameter information, reducing model size and power consumption, and accelerating performance
    - **Quantization error solution:**
        - Previous challenge: exponential quantization error accumulation in multi-layered neural networks (e.g., LLMs)
        - Fujitsu's solution: technology for quantization error propagation, a new quantization algorithm
        - Technology for quantization error propagation prevents error growth through cross-layer error propagation, based on theoretical insights
    - **1-Bit quantization achievement:**
        - 1-bit LLM quantization achieved via Fujitsu's proprietary, world-leading optimization algorithm for large-scale problems



2. **Specialized distillation for condensing expertise and improving accuracy:**
    - **Brain-inspired optimization:**
        - AI model structure optimization, mirroring brain processes of knowledge strengthening and memory organization
    - **Model generation & selection:**
        - Foundational AI model modification: pruning (unnecessary knowledge removal), transformer block additions (new capability impartation)
        - Diverse candidate model generation
        - Optimal model selection: Neural Architecture Search (NAS) with Fujitsu's proxy technology, balancing customer requirements (GPU resources, speed) and accuracy
    - **Knowledge distillation:**
        - Knowledge transfer from teacher models (e.g., Takane) into the selected structural model
    - **Beyond compression:**
        - Model compression with enhanced specialized task accuracy, surpassing foundational generative AI model performance

- **Demonstrated results (sales negotiation prediction):**
  - Text QA task (sales negotiation outcome prediction, Fujitsu CRM data):
    - 11x inference speed increase
    - 43% accuracy improvement
  - Student model (1/100th parameter size) achieved higher accuracy than teacher model
  - 70% reduction in GPU memory and operational costs
  - Enhanced sales negotiation outcome prediction reliability
- **Demonstrated results (image recognition):**
  - 10% improvement in unseen object detection accuracy (4) over existing distillation techniques
  - Significant breakthrough: over three times the accuracy improvement in this domain over two years



Future plans

Moving forward, Fujitsu will further enhance Takane leveraging this technology, empowering customer business transformations. Future plans include lightweight, specialized Takane-derived agentic AI for finance, manufacturing, healthcare, and retail. Further technology advancements aim for up to 1/1000 model memory reduction with maintained accuracy, enabling ubiquitous high-precision, high-speed generative AI. Ultimately, specialized Takane models will evolve into advanced AI agent architectures for deeper world understanding and autonomous complex problem-solving.

Notes:

1. **World's first:**

   The systematic combination of Neural Architecture Search (a technology that automatically explores the structure of AI by combining multiple algorithms) with knowledge distillation and its achievement of both lightweighting and accuracy that surpasses the original AI model.

2. **World's highest accuracy retention rate:**

   Fujitsu independent benchmark testing confirmed performance surpassing cutting-edge methods such as OneBit.

3. **Hugging Face:**

   A widely used platform for sharing and collaborating on machine learning models and datasets.
   https://huggingface.co/qep

4. Accepted at the 2025 IEEE International Conference on Image Processing (ICIP 2025).

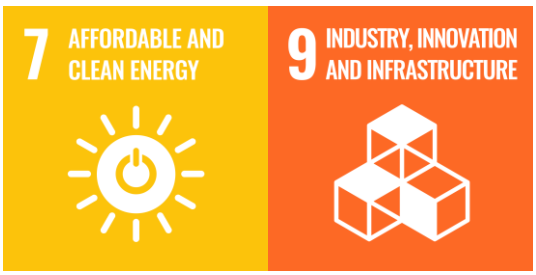
Related Links:

- Fujitsu Kozuchi
- Paper: "Quantization Error Propagation: Revisiting Layer-Wise Post-Training Quantization"
- Paper: "Optimization by Parallel Quasi-Quantum Annealing with Gradient-Based Sampling"

Fujitsu's Commitment to the Sustainable Development Goals (SDGs)

The Sustainable Development Goals (SDGs) adopted by the United Nations in 2015 represent a set of common goals to be achieved worldwide by 2030. Fujitsu's purpose — "to make the world more sustainable by building trust in society through innovation" — is a promise to contribute to the vision of a better future empowered by the SDGs.

The goals most relevant to this project



About Fujitsu

Fujitsu's purpose is to make the world more sustainable by building trust in society through innovation. As the digital transformation partner of choice for customers around the globe, our 113,000 employees work to resolve some of the greatest challenges facing humanity. Our range of services and solutions draw on five key technologies: AI, Computing, Networks, Data & Security, and Converging Technologies, which we bring together to deliver sustainability transformation. Fujitsu Limited (TSE:6702) reported consolidated revenues of 3.6 trillion yen (US$23 billion) for the fiscal year ended March 31, 2025 and remains the top digital services company in Japan by market share.
global.fujitsu

Press Contacts

Fujitsu Limited

Public and Investor Relations Division

Inquiries

Date: September 8, 2025

City: Kawasaki, Japan

Company: Fujitsu Limited