Kazuki Fujii

+81 80 9187 1150 - kazuki.fujii@rio.scrc.iir.isct.ac.jp
• linkedin.com/in/kazuki-fujii/ • github.com/okoge-kaz • scholar.google.co.jp/citations?user=jHXLs2wAAAAJ

Education

Tokyo Institute of Technology April 2020 — March 2024
Bachelor of Engineering in Computer Science

Institute of Science Tokyo (formerly Tokyo Institute of Technology) **April 2024 — March 2026** (expected)
Master of Engineering in Computer Science

Institute of Science Tokyo (formerly Tokyo Institute of Technology) **Admission Confirmed**/ Starting April 2026
Ph.D. in Computer Science

Experience

Swallow Project (**AIST:** National Institute of Advanced Industrial Science and Technology) **October 2023 – Present**
Core Contributor (Research Intern)
Tokyo, Japan

Systems Research & Large-Scale Training (2.5+ Years): Since 2023, I have been the **core engineer** responsible for selecting, maintaining, and extending core training libraries (**Megatron-LM**, **NeMo**). I independently handled all bug fixes, feature extensions, and upstream updates. Additionally, I architected experimental environments, including **cluster setups on AWS and Google Cloud,** and the configuration of environment modules for institutional supercomputers.

Execution of All Major Training Workloads (2.5+ Years): Orchestrated **every medium-to-large-scale pre-training and post-training run** for the project. This hands-on execution led to the release of nine bilingual LLM series. The first-generation models (Dec 2023) achieved SOTA among open Japanese LLMs, and the methodology paper has received **110+ citations**.

Dataset Refinement (SwallowCode **&** SwallowMath**) (1+ Years): Led** the research planning, experimental design, and data construction for both **v1 (May 2025)** and **v2 (Nov 2025)** releases. These datasets deliver some of the strongest performance on open code/math benchmarks. While I led the creation phase independently, I collaborated with the team for comprehensive LLM evaluations and visualizations.

Training Acceleration Research (1.5+ Years): To maximize training efficiency with limited compute resources, I spearheaded the validation of FP8 techniques for large-scale models. My investigations (June–Dec 2024) proved that standard Delayed Scaling and per-tensor CurrentScaling caused instability in **70B models**, challenging existing assumptions. Building on these findings, I validated the **Blockwise FP8 (E4M3)** method introduced by NVIDIA, ultimately establishing a stable training method that delivered a **1.18× FLOP/s/GPU speedup** with no loss in downstream accuracy.

Open Source & Collaboration: Authored llm-recipes **(44 stars)** and moe-recipes **(21 stars)** in 2023 to enable training for model architectures **unsupported by Megatron-LM at the time**. In parallel, I **collaborate with NVIDIA's Megatron-LM team** (PM Santosh Bhavani) to synchronize development roadmaps.

Preferred Networks
Research Engineer Intern
Tokyo, Japan **Jun 2025 – September 2025**

Member of the Plamo Pre-Training team, responsible for advancing Plamo, the company's proprietary large language model. I focused on improving trillion-token scale pre-training datasets and the practical adoption of low-precision training. Specifically, I leveraged code dataset refinement techniques I developed and publicly released in my academic lab, integrating them into Plamo's pre-training corpus and achieving more than a 10-point improvement on HumanEval+.

On the systems side, I conducted experiments with an 8B-parameter model to evaluate the stability of FP8 (E4M3) blockwise computation. Since full from-scratch training is prohibitively costly, I validated stability through continued pre-training from Llama-3.1-8B and Qwen3-8B over hundreds of billions of tokens. These experiments demonstrated that FP8 (blockwise) can deliver a 1.18× training speedup without causing instability or degradation in downstream task performance.

To reduce the high cost of training mid-sized(8B) LLMs every time new datasets need to be evaluated, I examined whether Bits Per Character (BPC)—which can be measured even with very small models—could be used to predict benchmark scores such as HumanEval and MMLU. The results showed that while there was correlation, BPC was not reliable enough to serve as a practical criterion for dataset adoption. This work was documented in Japanese, reviewed internally, and published as an official company technical blog post.

Turing

February 2023 – Present

Research Engineer Intern
Tokyo, Japan

VLM Infrastructure & Heron Development: Co-authored **vlm-recipes (20 stars)**, a PyTorch FSDP-based distributed training library, to address the lack of VLM support in Megatron-LM in **early 2024**. This library served as the core training infrastructure for the development of the **Turing VLM Heron** series.

Large-Scale Cluster Deployment: As part of the government-backed GENIAC program, collaborated with Google Cloud engineers to deploy **256 H100 GPUs (32 nodes)** using the **Cluster Toolkit**. As an early adopter, I managed the cluster provisioning and daily maintenance (including faulty node replacement) over a two-month intensive training period.

Environment Management: Configured and maintained environment modules across **H100 and H200 clusters**, ensuring stable development workflows for **50+ researchers and engineers**.

Thought Leadership & Public Speaking: Authored technical blog posts covering a wide range of topics—from **LLM training** to **large-scale infrastructure management**—that garnered **800+ likes**, significantly boosting the company's engineering presence. Notably, my insights on real-world Cluster Toolkit deployment led to an invitation to speak at **Google Cloud Next Tokyo 2024**, where I delivered a joint presentation with Google Cloud engineers.

SB Intuitions

April 2024 - May 2025

Research Internship
Tokyo, Japan

Core Library (Sarashina LLM): Led the complex migration of the company's proprietary Megatron-LM fork to the latest Megatron-Core. Successfully resolved extensive merge conflicts caused by custom research features (e.g., **embedding scaling**) and established a **custom CI pipeline** to verify stability and function correctness post-upgrade.

Checkpoint Conversion Engineering: Developed and validated internal checkpoint converters for **Llama-3 and Mixtral** prior to official support (pre-dating Megatron-Bridge). This critical tooling enabled seamless interoperability between Hugging Face and Megatron-LM formats for the research team.

MoE Infrastructure Implementation: Enabled internal Mixture-of-Experts (MoE) research by integrating my open-source library, **moe-recipes** (20+ stars), into the company's infrastructure in early 2024. I optimized the library for internal clusters and authored comprehensive documentation, significantly accelerating the adoption of MoE training by ML researchers.

Kotoba Technologies

October 2023 - February 2024

Researcher Intern
Tokyo, Japan

Library Development: As the **first employee**, I spearheaded the development of **kotoba-recipes (33 stars)** within two months of joining. This FSDP (v1)-based library provided a streamlined framework for instruction tuning Hugging Face models and later integrated Megatron-LM's dataloaders for efficient pre-training. It served as the **prototype for my widely used llm-recipes library**.

Pioneering Mamba Research: Rapidly responded to the release of the Mamba architecture (Dec 2023) by developing and open-sourcing **kotomamba (71 stars)** within a single month. I integrated Mamba into my training framework and conducted bilingual (Japanese/English) experiments, **validating for the first time** that Mamba architectures support language adaptation through continual pre-training, similar to Transformers.

Strategic Initiatives & Client Delivery: Collaborated directly with founders Jungo Kasai and Noriyuki Kojima to define the company's technical direction while delivering custom fine-tuned LLMs for external clients during the seed phase.

Technical Blog Contributions

NVIDIA Technical Blog Contributor: Authored a featured article on NVIDIA's official technical blog, sharing architectural insights and engineering challenges behind developing a **172B-parameter LLM**.

Technical Blog Writer: Authored **25 technical articles** on Zenn (Japan's leading engineering blog platform), accumulating **1,400+ likes**. Topics span corporate R&D case studies, academic research findings, and practical guides on distributed LLM training and infrastructure setup.

Awards

2024 Best Paper Award of IPSJ National Convention (Japan) (Distributed Training of LLMs)

2024 Best Paper Award of The Association for Natural Language Processing (Japan) (Continual Pre-Training of LLMs)

Research Papers

First Author Papers

Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities (COLM 2024)

112 citations, Swallow Project work

Cross-Lingual Adaptation & Impact: Developed **Swallow LLM** by continually pre-training from Llama 2, achieving SOTA performance among open Japanese LLMs (Dec 2023). Demonstrated the viability of adapting English-centric models to non-English languages through vocabulary extension and continual pre-training.

Scientific Insights: Conducted comprehensive ablation studies on data mixing, vocabulary expansion, and parallel corpora to optimize cross-lingual transfer.

Data Mixing: Identified that a **9:1 Japanese-English ratio** effectively mitigates the degradation of original English capabilities.

Parallel Corpora: Demonstrated that integrating **parallel corpora** significantly boosts translation performance without compromising general task capabilities.

Vocabulary: Revealed that vocabulary expansion improves generation efficiency by **78%** but introduces a specific trade-off by degrading summarization performance.

Rewriting Pre-Training Data Boosts LLM Performance in Math and Code (Under Review at ICLR 2026)

7 citations, Swallow Project work

Methodology (Transform-and-Retain): Proposed a novel **LLM-driven rewriting pipeline** for pre-training data that upgrades low-quality samples instead of discarding them. Introduced *Style-Guided* and *Self-Contained Optimization* rewriting for code dataset, and a context-restoration pipeline with step-by-step reasoning for math, distinguishing this approach from traditional exclusionary filtering or SFT data synthesis.

SOTA Performance: Demonstrated that training on refined data boosts Llama-3.1-8B performance significantly within a fixed token budget. **SwallowCode** outperformed Stack-Edu by **+17.0 pass@1 on HumanEval** , and **SwallowMath** surpassed finemath-4+ by **+12.4 accuracy on GSM8K**.

Open Science & Impact: Released **SwallowCode** and **SwallowMath** (v1) and the data processing pipeline to the community. Updated the datasets to v2 with an improved pipeline and released them under an **Apache-2.0 license** to facilitate broader adoption.

Accelerating Large Language Model Training with 4D Parallelism and Memory Consumption Estimator (2024)

3 citations

Theoretical Modeling: Formulated a precise memory consumption equation for **4D parallelism** (Data, Tensor, Pipeline, and Context Parallelism) in Llama architectures. The equation accounts for often-neglected factors such as activation recomputation and memory fragmentation, thereby surpassing existing GPT-centric estimators.

Empirical Analysis: Conducted **454 large-scale experiments** across NVIDIA A100 and H100 clusters to validate the estimator. Established a robust empirical threshold, demonstrating that training is guaranteed to be OOM-free when estimated usage is below **80%** of GPU memory.

Optimization Impact: Enabled the preemptive identification of optimal distributed training configurations without requiring actual GPU execution. This methodology significantly reduces the search space for optimal hyperparameters when migrating to new hardware generations or cluster topologies.

Balancing speed and stability: The trade-offs of fp8 vs. bf16 training in llms (2024)

3 citations, (FP8 GEMM, FP8 DelayedScaling, Loss Spike)

Co-Author Papers (selected)

Building a Large Japanese Web Corpus for Large Language Models (COLM 2024)

25 citations, Swallow Project work (dataset, continual pre-training, filtering)

Drop-Upcycling: Training Sparse Mixture of Experts with Partial Re-initialization (ICLR 2025)

6 citations (MoE, Upcycling)

Building Instruction-Tuning Datasets from Human-Written Instructions with Open-Weight Large Language Models (COLM 2025)

2 citations, Swallow Project work (instruction tuning dataset)

Skills

Languages: Python, C++, CUDA, Framework: PyTorch, Megatron-LM, NeMo

Research: LLM Pre-Training/Post-Training(SFT, RL), Distributed Training(TP, CP, EP, PP, DP), FP8, VLM, Dataset Refinement