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

Institute of Science Tokyo **Expected: March 2026**
Master of Engineering in Computer Science

- Specialization: High-Performance Computing (HPC), Distributed Training, Low-Precision Training (FP8, Blockwise FP8).
- Research focus: Sovereign LLM development in Japan. (Swallow Project)

Experience

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

I am involved in selecting and maintaining pre-training and post-training libraries, managing experiments, and setting up experimental environments to develop Japanese LLMs with competitive performance. This initiative, known as the Swallow Project, has significantly advanced non-English LLM development, producing nine bilingual Japanese–English LLM series through continual pre-training. The first-generation Swallow models, released in December 2023, achieved state-of-the-art performance among open Japanese LLMs, and the paper describing their training methodology has since received 90+ citations.

Beyond model training, I constructed billion-scale pre-training corpora (SwallowCode and SwallowMath), which, as of May 2025, deliver the strongest performance for open code and open math datasets worldwide.

Since 2024, I have also led research on FP8-based training acceleration, demonstrating that FP8 current scaling caused instability and degraded downstream task performance, whereas FP8 (E4M3) blockwise achieved a 1.18× improvement in FLOP/s per GPU with no loss in downstream accuracy.

From October 2023 onward, I have independently executed a wide scope of tasks: continual pre-training, instruction tuning, preparation of math and code datasets, and the maintenance and feature extension of both Megatron-LM and our open-source instruction tuning library(llm-recipes: 42 GitHub stars). In parallel, I collaborate with NVIDIA Megatron-LM team, regularly aligning with Project Manager Santosh Bhavani on the Megatron-LM development roadmap to synchronize our project's LLM training strategies.

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

Member of the Plamo Pre-Training team, responsible for advancing Plamo, the company's proprietary large language model. I focus 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

I developed vlm-recipes, a PyTorch FSDP (v1)-based distributed training library for vision–language models that earned 20 GitHub stars. Since Megatron-LM did not support VLM training until the summer of 2024, I developed this new library together with another intern, enabling distributed training for vision–language models.

I also configured and maintained environment modules across two clusters (H100 and H200) used by 50+ researchers and engineers, and—as part of Japan's government-backed GENIAC program—collaborated with Google Cloud engineers to utilize the Cluster Toolkit library for deployment, successfully bringing up 32×H100 nodes (=256 GPUs) and handling daily maintenance (faulty GPU node replacement) over two months. The knowledge gained from this work was published in Japanese as a technical blog. At that time, real-world use cases of Cluster Toolkit were still rare, which led to an invitation to speak at Google Cloud Next Tokyo 2024, where I gave a joint presentation together with Google Cloud engineers. In addition, my posts on the company's official Tech Blog received over 700 likes, contributing both to the company's visibility and to the broader community.

SB Intuitions **April 2024 - May 2025**
Research Internship
Tokyo, Japan

As a member of the Pre-Training team for Sarashina, the company's proprietary large language model, I worked on maintaining the pre-training library (a fork of Megatron-LM) and improving the pre-training dataset. Our in-house Megatron-LM fork contained several features originally proposed by the Sarashina team in research papers, such as coefficient-based embedding scaling. Upgrading to a newer version of Megatron-Core resulted in numerous conflicts, which I resolved, and I subsequently built a custom CI pipeline to verify that the library remained functional after the upgrade.

During my internship (2024–spring 2025), Megatron-Bridge didn't exist to convert between Hugging Face and Megatron checkpoints. Before the official checkpoint converter supported Llama-3 and Mixtral, I implemented an internal converter and validated its correctness. In addition, because Megatron-LM's support for Mixture-of-Experts (MoE) training was still limited in early 2024, I adapted my own open-source MoE training library, moe-recipes (20+ stars), for use within the company. This included ensuring its proper operation on internal clusters and preparing documentation so that ML researchers could use it effectively.

Kotoba Technologies **October 2023 - February 2024**
Researcher
Tokyo, Japan

I joined the company as its first employee outside of the founders. At a time when the company's direction was still being defined, I collaborated with the founders, Jungo Kasai and Noriyuki Kojima, on various commissioned projects, developing fine-tuned LLMs for external clients, while also discussing the next steps for the company.

During this period, I almost single-handedly developed the prototype of what would later become llm-recipes: a library called kotoba-recipes (33 stars), which I released as open source only two months after joining. This library aimed to provide a simple framework for instruction tuning any Hugging Face Transformers-based LLM, and I adopted FSDP v1 as the distributed backend. Later, to support continued pre-training, I incorporated Megatron-LM's dataset class and dataloader implementations, enabling efficient data loading.

In addition, we focused on Mamba, which had just been released in December 2023. By forking Tri Dao's implementation and integrating it with kotoba-recipes, I developed kotomamba (71 stars) within a month and subsequently open-sourced it. Using kotomamba, I conducted continued pre-training in both Japanese and English and demonstrated that, similar to Transformer models, Mamba also supports language adaptation through continual pre-training.

Technical Blog Contributions

I have authored 23 technical blog articles on the Japanese technical blog platform Zenn, which have collectively received 1,301 likes. These include articles representing both corporate and academic projects, as well as my personal technical blogs. The topics span a wide range, including explanations of LLM training methodologies, practical tips from research and development, and short technical reports on LLM development. In addition, I contributed an article to NVIDIA's official Technical Blog, sharing my experience in developing a 172B-parameter LLM.

Research Papers

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

98 citations, First Author

- Conducted continual pre-training of the Swallow LLM, enhancing Japanese capabilities by extending Llama 2's vocabulary and training on a 100B-token Japanese web corpus, achieving **SoTA performance on Japanese** tasks at December 2023.
- Single-handedly prepared and **maintained pre-training and post-training libraries**, optimizing workflows for efficient cross-lingual model development and experimentation.
- Achieved 95% GPU utilization during a one-month large-scale training experiment on 50 nodes (400 A100 GPUs) through careful planning and dedicated effort.

Paper: Rewriting Pre-Training Data Boosts LLM Performance in Math and Code **(Under Reviewing ICLR 2026)**

6 citations, First Author

- Introduced SwallowCode and SwallowMath, two openly licensed datasets, enhancing LLM performance in program synthesis and mathematical reasoning through a novel transform-and-retain pipeline that refines public data.
- Demonstrated significant performance improvements in Llama-3.1-8B, achieving +17.0 pass@1 on HumanEval and +12.4 accuracy on GSM8K within a 50B token budget.

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

2 citations, First Author

- Developed precise memory consumption formulas for 4D parallel training (DP, TP, PP, CP) in the Llama architecture, enabling preemptive identification of memory overflow and significantly reducing the configuration search space.
- Provided empirical insights into optimal 4D parallelism configurations through comprehensive analysis of 454 experiments on A100 and H100 GPUs.

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

21 citations, Co-Author

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

4 citations, Co-Author

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

1 citation, Co-Author

Awards

2024 **Best Paper Award** of IPSJ National Convention (Japan)

2024 Best Paper Award of **The Association for Natural Language Processing (Japan)**