taishi@rio.scrc.iir.isct.ac.jp
taishi-n324.github.io
taishi-n324
taishi-nakamura
Setuna7777_2
nbPQwgUAAAAJ

# Taishi Nakamura

## Education

04/2024–03/2026 (expected) **MS in Computer Science**, *Institute of Science Tokyo*, Tokyo, Japan
Supervisor: Rio Yokota

04/2021–03/2024 **BS in Computer Science**, *Tokyo Institute of Technology*, Tokyo, Japan
Completed degree one year early in recognition of academic excellence

## Employment

05/2024–present **Research Assistant**, *National Institute of Informatics Research and Development Center for Large Language Models*
Working with Prof. Jun Suzuki and Prof. Rio Yokota

02/2024–present **Research Intern**, *Sakana AI*, Tokyo, Japan
Working with Dr. Takuya Akiba

08/2023–present **Student Trainee**, *National Institute of Advanced Industrial Science and Technology*
Working with Dr. Hiroya Takamura,
Prof. Naoaki Okazaki, Prof. Rio Yokota

04/2023–present **Research Assistant**, *Institute of Science Tokyo*, Tokyo, Japan
Working with Prof. Rio Yokota

10/2023–04/2024 **Research Intern**, *LLM-jp Project*
Working with Prof. Jun Suzuki

06/2023–02/2024 **MITOU Target Program Fellow**, *Information-technology Promotion Agency*
Mentor: Dr. Yuuki Tokunaga

## Research Interests

I am interested in efficient training of large language models, including scaling and training Mixture-of-Experts (MoE) models, continual pre-training, and efficient post-training with reinforcement learning algorithms.

## Selected Awards, Grant

2025 **Language Resources Award at Annual Meeting of the Japanese Association for NLP**

2024 **Outstanding Paper Award at Annual Meeting of the Japanese Association for NLP**

2023 **Keyence Foundation 2023 'Go for it! Japanese University Students' Support Grant**

## Projects

**LLM-jp** – Core Contributor LLM-jp is a fully open, scratch-trained series of large language models primarily focused on Japanese. The project emphasizes transparency in data, training, and model design. Contributed to the design and large-scale training of the LLM-jp models. For LLM-jp-3, applied a novel sparse Mixture-of-Experts training strategy (Drop-Upcycling; ICLR 2025) to scale the model to 8×13B parameters on 256 GPUs, achieving superior performance compared to a 172B dense model trained on the same data. At the time, it was among the most competitive open Japanese LLMs. For LLM-jp-4, designed the model architecture and training strategy based on findings from our ICLR 2026 work on optimal sparsity in MoE language models.

**Swallow LLM** – Swallow LLM is a series of Japanese-specialized large language models developed through continual pre-training. Co-led the early development of Swallow LLM, establishing effective continual pre-training methods for Japanese LLM adaptation (COLM 2024) and designing synthetic instruction-tuning data generation pipelines for early-stage alignment. Led large-scale continual pre-training and post-training of Gemma-2 Swallow on TPU v6e (256 chips), achieving state-of-the-art Japanese performance in the 2B and 9B model classes. The 27B model demonstrated performance comparable to 70B-scale models. Recently led reinforcement learning-based post-training for Qwen3-Swallow and GPT-OSS Swallow to further improve reasoning capabilities.

**Aurora-M (COLING 2025)** – Led Aurora-M, an open-source multilingual and code-focused continual pre-training project. Coordinated cross-institutional collaboration and executed large-scale distributed training on the Lumi supercomputer (32 nodes, AMD GPUs). Contributed to open-science initiatives for multilingual and code LLM development.

## Publications

### Peer-Reviewed Conference Publications

Kazuki Fujii, Yukito Tajima, Sakae Mizuki, Masaki Kawamura, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Masanari Oi, **Taishi Nakamura**, Takumi Okamoto, Shigeki Ishida, Kakeru Hattori, Youmi Ma, Hiroya Takamura, Rio Yokota, Jun Sakuma, and Naoaki Okazaki. Rewriting pre-training data boosts LLM performance in math and code. In *International Conference on Learning Representations (ICLR)*, 2026.

**Taishi Nakamura**, Satoki Ishikawa, Masaki Kawamura, Takumi Okamoto, Daisuke Nohara, Jun Suzuki, and Rio Yokota. Optimal sparsity of mixture-of-experts language models for reasoning tasks. In *International Conference on Learning Representations (ICLR)*, 2026. **Oral**.

So Kuroki, **Taishi Nakamura**, Takuya Akiba, and Yujin Tang. Agent skill acquisi-

tion for large language models via cycleQD. In *International Conference on Learning Representations (ICLR)*, 2025.

Youmi Ma, Sakae Mizuki, Kazuki Fujii, **Taishi Nakamura**, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, et al. Building instruction-tuning datasets from human-written instructions with open-weight large language models. In *Conference on Language Modeling (COLM)*, 2025.

**Taishi, Nakamura***, Mayank, Mishra*, Simone, Tedeschi*, Chai Yekun, Stillerman Jason T, Friedrich Felix, Yadav Prateek, Laud Tanmay, Chien Vu Minh, Zhuo Terry Yue, et al. Aurora-M: Open source continual pre-training for multilingual language and code. In *International Conference on Computational Linguistics (COLING): Industry Track*, 2025.

**Taishi Nakamura**, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. Drop-upcycling: Training sparse mixture of experts with partial re-initialization. In *International Conference on Learning Representations (ICLR)*, 2025.

Yuichi Inoue*, Kou Misaki*, Yuki Imajuku, So Kuroki, **Taishi Nakamura**, and Takuya Akiba. Wider or deeper? scaling LLM inference-time compute with adaptive branching tree search. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2025. **Spotlight**.

Kazuki Fujii*, **Taishi Nakamura***, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *Conference on Language Modeling (COLM)*, 2024. **Outstanding Paper & Language Resource Award at Annual Meeting of the Japanese Association for NLP**.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, **Taishi Nakamura**, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In *Conference on Language Modeling (COLM)*, 2024. **Outstanding Paper Award at Annual Meeting of the Japanese Association for NLP**.

### Workshop Publications

Marianna Nezhurina*, Jörg K.H. Franke, **Taishi Nakamura**, Timur Carstensen, Niccolò Ajroldi, Ville Komulainen, David Salinas, and Jenia Jitsev*. open-sci-ref-0.01: open and reproducible reference baselines for language model and dataset comparison. In *AAAI Workshop on Reproducible AI*, 2025.

Koshiro Saito, Sakae Mizuki, Masanari Ohi, **Taishi Nakamura**, Taihei Shiotani, Koki Maeda, Youmi Ma, Kakeru Hattori, Kazuki Fujii, Takumi Okamoto, et al. Why we build local large language models: An observational analysis from 35 japanese and multilingual LLMs. In *Conference on Language Modeling (COLM) Multilingual and Equitable Lan-guage Technologies Workshop*, 2025. **Outstanding Paper Award, Natural Language Processing Research Meeting of the Information Processing Society of Japan**.

Kazuki Fujii, **Taishi Nakamura**, and Rio Yokota. llm-recipes: A framework for seamless integration and efficient continual pre-training of large language models. In *International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) Trillion Parameter Consortium Workshop*, 2024.

### Preprints

Huu Nguyen*, Victor May*, Harsh Raj*, Marianna Nezhurina, Yishan Wang, Yanqi Luo, Minh Chien Vu, **Taishi Nakamura**, Ken Tsui, Van Khue Nguyen, David Salinas, Aleksandra Krasnodębska, Christoph Schuhmann, Mats Leon Richter, Xuan-Son, Vu, and Jenia Jitsev. Mixturevitae: Open web-scale pretraining dataset with high quality instruction and reasoning data built from permissive-first text sources. arXiv:2509.25531, 2025.

Kazuki Fujii, **Taishi Nakamura**, and Rio Yokota. Balancing speed and stability: The trade-offs of fp8 vs. bf16 training in llms. arXiv:2411.08719, 2024.

LLM-jp. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. arXiv:2407.03963, 2024.

## Service

### Reviewer

2025–present **ICLR 2026**

2025–present **Workshops**, *ICML 2025 AI4Math, NeurIPS 2025 ER*