Shota Suzuki

Ph.D. Student in Electrical Engineering and Information Systems | The University of Tokyo
Email (via Web) | Portfolio Website | Google Scholar | LinkedIn

RESEARCH INTERESTS & EXPERTISE

I am conducting research on novel memory-centric computing systems, tackling hardware-software co-design challenges by leveraging knowledge across algorithms, computer architecture, circuits, and semiconductor devices.

Areas of Expertise: Large Language Models (LLMs), Computer Architecture, Computation-in-Memory (CiM), Resistive Random-Access Memory (ReRAM)

EDUCATION

- **The University of Tokyo** | Tokyo, Japan

Ph.D. in Electrical Engineering and Information Systems (Apr 2026 – Present)

  - Advisor: Prof. Ken Takeuchi
  - Project: Advanced Human Resource Development Leading Green Transformation (SPRING GX)

- **The University of Tokyo** | Tokyo, Japan

Master of Engineering in Electrical Engineering and Information Systems (Apr 2024 – Mar 2026)

  - Advisor: Prof. Ken Takeuchi
  - Master's Thesis: *"Research on Analysis of Changes in Current by Read-disturb on Multi-level Cell ReRAM and ReRAM Application for KV Cache of Large Language Models"*

- **The University of Tokyo** | Tokyo, Japan

Bachelor of Engineering in Electrical and Electronic Engineering (Apr 2020 – Mar 2024)

  - Advisors: Prof. Shuichi Sakai and Prof. Hidetsugu Irie
  - Graduation Thesis: *"Microarchitecture Design of Dynamic Approximate Architecture Based on Execution Time"*

GRANTS & SCHOLARSHIPS

- **Merit-Based Scholarship Repayment Exemption (Half-Exemption)**, JASSO (Jul 2026)
  - Awarded for outstanding academic and research achievements during the master's program.
- **Supporting Pioneering Research through AI for 1,000 Discovery challenges Program (SPReAD)**, MEXT (Jul 2026 – Jan 2027)
  - Research Project: "Creation of Highly Power-Efficient AI Hardware: Agent-Driven Autonomous Processing-in-Memory Design" (Original Jp: 高電力効率な AI ハードウェア創出：エージェント主導の Processing-in-Memory 自律設計)
  - Role: Principal Investigator (PI)
  - Awarded a highly competitive research grant aimed at driving scientific innovation through advanced AI technologies.
- **Research Fellowship / Grant (SPRING GX)**, JST SPRING Program, The University of Tokyo (Apr 2026 – Mar 2029)
- **Student Supporters Club Scholarship (Grant-type)**, The University of Tokyo Foundation (Apr 2024 – Mar 2026)

AWARDS

- **SLDM Excellent Presentation Award for FY2025**, Special Interest Group on System and LSI Design Methodology (SLDM) of Information Processing Society of Japan (IPSJ) (Aug 2026, Scheduled)
- **Qualcomm Award (Special Prize)**, *UTokyo AI Program Hackathon*, Qualcomm Japan, Microsoft Japan, and Graduate School of Engineering, The University of Tokyo (Aug 2026)
    - Recognized for developing a local AI health assessment app on Snapdragon X Surface Pro using webcam-based rPPG and MediaPipe.
    - Served as team leader (project planning/management) and lead developer for rPPG tool selection and data integration into local AI models.
    - Acquired technical expertise in Windows ARM64 Python builds and Qualcomm NPU acceleration.
- **1st Place / Best Project Award**, *Web Engineering & Business Model Course*, Matsuo-Iwasawa Lab, The University of Tokyo (Jul 2026)
    - Awarded 1st place among 4 competing teams for developing the concept of “DIY Blueprint Maker,” an application that automatically generates easy-to-follow DIY design plans for complete beginners.
    - Initiated the product vision and served as a domain advisor, applying practical experience in theater stage set design to guide structural feasibility and the construction process for users.
- **Award in the Large Language Model (LLM) Category**, *Chipathon (Semiconductor Design Hackathon)*, Graduate School of Engineering, The University of Tokyo (Feb 2026)
    - Developed and optimized a hardware-accelerated system to execute a language model on an FPGA board.
- **Best Research Award**, *The 8th cross-disciplinary Workshop on Computing Systems, Infrastructures, and Programming (xSIG 2024)* (Aug 2024)

RESEARCH & WORK EXPERIENCE

- **TOWN, Inc.** | *Part-time Engineer* (Oct 2022 – Sep 2025)
    - Built and operated cloud infrastructure using Amazon Web Services (AWS).
    - Automated infrastructure testing using Serverspec.
    - Participated in an R&D project to automate server configuration and construction using Microsoft Azure OpenAI.
- **University of Leoben** | Leoben, Austria
  Research Intern (IAESTE Internship Program) (Aug 2024 – Sep 2024)
    - Conducted research on the analysis of X-ray diffraction spectra using neural networks under the mentorship of Assoc. Prof. Ernst Gamsjäger.
- **The University of Tokyo** | *Teaching Assistant (TA)* (Mar 2022 – Jul 2025)
  Assisted and coordinated multiple laboratory and undergraduate courses:
    - *“Prototyping Experience to Shape Ideas - Prototype Design in the AI Era”* (Mar 2025 – Jul 2025)
    - *“Approximate Computing for Machine Learning Applications”* (Oct 2024 – Dec 2024)
    - *“Prototyping Experience to Shape Ideas - From Robots to Home Appliances”* (Multiple terms, served as TA Coordinator in 2023)
    - *“Analog Circuits”* (May 2024)
    - *“VLSI Architecture”* (Mar 2024 – May 2024)

CERTIFICATES & PROFESSIONAL TRAINING

- **Deep Learning Advanced Course: Deep Generative Models**, Matsuo-Iwasawa Lab, The University of Tokyo (Apr 2026)
  Advanced practical program on theories and implementation of VAEs, GANs, Energy-Based Models, and Diffusion Models.
- **AI and Semiconductor Course**, Systems Design Lab & Matsuo-Iwasawa Lab, The University of Tokyo (Oct 2025)
  Comprehensive training in semiconductor technologies, architecture, and hardware/software co-design for AI workloads.
- **Deep Learning Basic Course**, Matsuo-Iwasawa Lab, The University of Tokyo (Jul 2025)
  Practical program on fundamental deep learning theories and frameworks (PyTorch, CNNs).
- **Global Consumer Intelligence (GCI) Program**, Dept. of Technology Management for Innovation, The University of Tokyo (Feb 2024)
  Data science and machine learning training (data analysis and predictive modeling using Python).
- **Go Global Gateway (GGG) Certificate of Recognition**, Center for Global Education, The University of Tokyo (Nov 2023)
  Recognized for international engagement (experiential learning at KTH Sweden, study abroad at University of Hawaii, and UTokyo GUC Student Ambassador).
- **Participant, Boeing Externship Program 2023**, Boeing Japan and Division of Engineering, Institute for Innovation in International Engineering Education, The University of Tokyo
  Participated in a practical project development and proposal presentation related to aerospace systems under the guidance of Boeing Japan.

PUBLICATIONS

Refereed Journal Articles

1. Chihiro Matsui, Ayumu Yamada, **Shota Suzuki**, Naoko Misawa, and Ken Takeuchi, “NanoCiM: Nano-ReRAM Computation-in-Memory with Neural Network Weight-aware X-Address Scrambling to Suppress IR Drop-oriented Bit-line MAC Current Variation,” *IEEE Transactions on Circuits and Systems for Artificial Intelligence (TCASAI)*, vol. 3, no. 3, pp. 190-202, June 2026.
2. **Shota Suzuki**, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “Mapping strategy of quantized KV cache of LLMs into mixture of SLC and MLC ReRAM among various quantization methods,” *Japanese Journal of Applied Physics (JJAP)*, vol. 65, pp. 03SP09, February 9, 2026.
3. Yuya Degawa, **Shota Suzuki**, Junichiro Kadomoto, Hidetsugu Irie, and Shuichi Sakai, “Cycle-Oriented Dynamic Approximation: Architectural Framework to Meet Performance Requirements,” *IEEE Computer Architecture Letters (CAL)*, vol. 23, no. 2, pp. 211-214, July-December 2024.

Refereed International Conferences

1. Koki Shibata, **Shota Suzuki**, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “CRISP: Partitioning LLM Inference Across Main Memory and HBM via Input Sparsification and Low-Rank Approximation,” *The 8th IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS 2026)*, (Accepted).
2. Masafumi Higashi, **Shota Suzuki**, Yuya Yamamoto, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “KV Cache Management for Block Diffusion Language Models on Multi-Level NV-Memory,” *The 8th IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS 2026)*, (Accepted).

3. **Shota Suzuki**, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “Hybrid SLC/MLC ReRAM for Quantized KV Cache in Transformer-based LLM,” *International Conference on Solid State Devices and Materials (SSDM)* Poster Late News, September 17, 2025, pp. 755-756.
4. **Shota Suzuki**, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “Observation of State Change in 40nm TaO_X ReRAM Cells during Read-disturb,” *IEEE Silicon Nanoelectronics Workshop (SNW)*, June 8, 2025, pp. 26-27.

Refereed Domestic Conferences

1. **Shota Suzuki**, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “Hybrid SLC/MLC ReRAM Mapping Scheme for Quantized KV Cache of LLMs Across Various Quantization Methods,” (Original Jp: 多様な量子化手法に対応した LLM 量子化 KV キャッシュの SLC/MLC ハイブリッド ReRAM マッピング手法), *DA Symposium 2026 -System and LSI Design Methodology-*, (Scheduled), **(SLDM Excellent Presentation Award for FY2025)**.
2. **Shota Suzuki**, Yuya Degawa, Tomohiro Yoshita, Junichiro Kadomoto, Hidetsugu Irie, and Shuichi Sakai, “RTL Design and Soft Processor Implementation of an Architecture for Cycle-oriented Dynamic Approximation,” *The 8th cross-disciplinary Workshop on Computing Systems, Infrastructures, and Programming (xSIG 2024)*, August 7, 2024, pp. 1-14. **(Best Research Award)**

Non-Refereed Domestic Conferences / Workshops

1. **Shota Suzuki**, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “Investigation of Current Fluctuations and RTN Frequency Change Induced by Read Stress in ReRAM,” *The 73rd JSAP Spring Meeting*, March 15, 2026.
2. **Shota Suzuki**, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “ReRAM Mapping Method Based on Bit-Level Error Robustness of Quantized KV Cache,” *The 73rd JSAP Spring Meeting*, March 15, 2026.
3. **Shota Suzuki**, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “Investigation of ReRAM Mapping Techniques for Storing Quantized KV Cache in LLMs,” *IEICE Technical Report*, vol. 125, no. 260, VLD2025-59, pp. 227-232, December 3, 2025. **(SLDM Excellent Presentation Award for FY2025)**
4. **Shota Suzuki**, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi, “Changes in Read Current and RTN Frequency due to Read-disturb in 40nm TaO_X ReRAM,” *IEICE Technical Report*, vol. 125, no. 261, ICD2025-50, pp. 74-79, December 2, 2025.
5. Daichi Mikubo, **Shota Suzuki**, Shu Sugita, Junichiro Kadomoto, and Hidetsugu Irie, “Acceleration of speech feature extraction using an architecture with dynamically controllable computational precision,” *The 24th Forum on Information Technology (FIT 2025)*, September 4, 2025.
6. Yifan Wang, Adil Padiyal, Daqi Lin, **Shota Suzuki**, Chihiro Matsui, and Ken Takeuchi, “A Design for Digital-CiM INT8 Transformer Accelerator with Pipeline,” *The 72nd JSAP Spring Meeting*, March 14, 2025.

SKILLS & INTERESTS

- **Languages:** Japanese (Native), English (Proficient), German (Duolingo 900+ days streak)
- **Interests & Hobbies:** Alpine Skiing (SAJ Level 2), Language Learning