Ruijie Zhu

Github | LinkedIn | Google Scholar | Personal Website |

Education

2019 - 2023 B.S. at **University of Electronic Science and Technology of China** (GPA: 3.83/4.0)
2023 - 2026 Ph.D. at **University of California, Santa Cruz** Expected Graduation: 2026.6

Experience

Student Researcher Intern

Jan 2025 - Dec 2025

Bytedance Seed

At ByteDance Seed I led two pre-training lines: data synthesis via Massive Genre Audience that produced a 770 billion token MGACorpus, **demonstrating that synthetic data generated by smaller models can effectively enhance the performance of larger models**; and latent reasoning pre-training with the Ouro Looped Language Model that uses iterative latent computation and entropy-regularized depth and scales to 7.7 trillion tokens. During the internship I used 1000 H200s to train the world's first industrial-grade Looped Language Model, owning the entire pipeline end to end from data curation, pre-training, mid-training, supervised fine-tuning to post-training reinforcement learning. I served as the primary pre-training and mid-training lead and also owned the inference engine and RL stack, and **our 1.4B and 2.6B models match the performance of Qwen3 models that are four times larger.** Through this experience I gained hands-on understanding of the full-cycle LLM training process at scale.

Selected Projects

Scaling Looped Transformers at Industrial Scale

Paper

Nov 2025

- **Led the entire pre-training infrastructure**, building the training stack from scratch and scaling it to **1,000 GPUs**.
- Owned the full data pipeline, curating and training on **7.7 trillion tokens**.
- Built the vLLM-based inference infrastructure and resolved the key reinforcement-learning and inference bottlenecks.
- Delivered **the world's first industrial-grade Looped Language Model trained from scratch end to end**, where the 1B model matches Qwen3-4B and the 2B model matches Gemma3-12B.

Scaling Concept LLM

Paper

Jan 2026

- Designed the entire concept-model design pipeline, implementing a training pipeline purpose-built for concept models with **Flex Attention** and **Flash Attention Varlen**.
- **Made the key architectural decisions that turned a non-converging model into a converging one.**

MGA: Massive Genre–Audience Data Synthesis

Paper

Jun 2025

- **Led the data pipeline** that achieved large-scale pre-training data augmentation.
- **Demonstrated that synthetic data generated by smaller models can effectively improve the learning of much larger models.**

Linear-Attention Distillation & the First Pure-RNN LLM GSA RWKV

Mar 2024

- **Led the distillation of full (softmax) attention into linear attention**, transferring knowledge over 100B tokens and **cutting pre-training cost by 20–30**× by preserving softmax.
- Designed the long-context pre-training infrastructure for linear-attention models (2024).
- Through RWKV, **scaled the world's first pure-RNN language model** to performance on par with the contemporary SOTA.

MatMul-free LM & Neuromorphic Deployment MatMul-free Loihi 2

Jun 2024

- Wrote the first fused Triton kernel combining **BitNet and LayerNorm**, **speeding up end-to-end training by over 25%**.
- **Led the entire project** and deployed the model on Intel's **Loihi 2** neuromorphic hardware.
- Achieved **100 tokens/s at 2W** on Loihi 2 — nearly **5× more energy-efficient than NVIDIA Jetson** (NVIDIA's most power-efficient platform).

Selected Publications

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, et al. (2025). "Scaling latent reasoning via looped language models". In: *arXiv preprint arXiv:2510.25741*.

Rui-Jie Zhu, Qihang Zhao, and Jason K Eshraghian (2024). "SpikeGPT: Generative Pre-trained Language Model with Spiking Neural Networks". In: *Transactions on Machine Learning Research (TMLR)*.

Rui-Jie Zhu, Yu Zhang, Ethan Sifferman, Tyler Sheaves, Yiqiao Wang, Dustin Richmond, Peng Zhou, and Jason K Eshraghian (2024). "Scalable MatMul-free Language Modeling". In: *arXiv:2406.02528*.

Xintong Hao, **Rui-Jie Zhu**, Ge Zhang, Ke Shen, and Chenggang Li (2026). "Reformulation for Pretraining Data Augmentation". In: *The Fourteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=dIOYpj9K8P.

Rui-Jie Zhu, Ziqing Wang, Leilani Gilpin, and Jason Eshraghian (2024). "Autonomous driving with spiking neural networks". In: *Advances in Neural Information Processing Systems* 37, pp. 136782–136804.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, et al., and **Rui-Jie Zhu** (2023). "RWKV: Reinventing RNNs for the Transformer Era". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 23)*.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, et al., and **Rui-Jie Zhu** (2024). "Eagle and Finch: RWKV with matrix-valued states and dynamic recurrence". In: *Conference on Lanugage Modeling (COLM)*.

Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, **Rui-Jie Zhu**, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo (2025). "Magictime: Time-lapse video generation models as metamorphic simulators". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuhong Chou, Man Yao, Kexin Wang, Yuqi Pan, **Rui-Jie Zhu**, Jibin Wu, Yiran Zhong, Yu Qiao, Bo XU, and Guoqi Li (2024). "MetaLA: Unified Optimal Linear Approximation to Softmax Attention Map". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=Y8YVCOMEpz.

Yu Zhang, Songlin Yang, **Rui-Jie Zhu**, Yue Zhang, Leyang Cui, Yiqiao Wang, Bolun Wang, Freda Shi, Bailin Wang, Wei Bi, Peng Zhou, and Guohong Fu (2024). "Gated Slot Attention for Efficient Linear-Time Sequence Modeling". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=jY4PhQibmg.

Steven Abreu, Sumit Bam Shrestha, **Rui-Jie Zhu**, and Jason K Eshraghian (2025). "Neuromorphic Principles for Efficient Large Language Models on Intel Loihi 2". In: *arXiv preprint arXiv:2503.18002*.