Zhenyu "Allen" Zhang

Member of Technical Staff, SpaceXSI (xAI) — Ph.D., The University of Texas at Austin

Palo Alto, CA, USA | Email | Website | LinkedIn | Google Scholar | X | GitHub

Research Interests

My research aims to build autonomous, self-improving AI systems that are verifiable and useful to humans. Toward this goal, I have worked on several layers of foundation-model development:

- *Data:* synthetic-data flywheels for post-training **Grok 4.5** and **Grok 4.6**.
- *Algorithms:* memory-efficient optimizers for large-scale training (**APOLLO** [1], **GaLore** [2]).
- *Infrastructure:* KV-cache compression for efficient model serving (**H_2O** [6]).
- *Evaluation:* benchmarks that measure how far AI can improve AI (**Singularity-Bench** [10]).

Awards and Honors

Thinking About Thinking Fellowship	2026
CPAL Rising Star Award, CPAL	2026
Outstanding Paper Honorable Mention, MLSys (*i.e.*, Best Paper Finalist)	2025
ML and Systems Rising Star, MLCommons	2025

Professional Experience

SpaceXSI (xAI), Palo Alto, CA — Mar 2026 – present

Member of Technical Staff, Post-training team: Long-horizon RL environment.

- **Grok 4.3**: Contributed to safety evaluation.
- **Grok 4.5**: Core contributor to tool-use capability for office work. **Debuted at #1** on Artificial Analysis's AutomationBench-AA (51.4%, up from 1.4% for Grok 4.3) and τ^3-Banking (33%, up from 12.4%).
- **Grok 4.6**: Co-led AI R&D capability, covering kernel optimization and inference infrastructure. **Led all tested models** on SpaceXAI MTS Eval (61.1%, vs. 40.6% for Grok 4.3) and InferenceEval (46.9%); three of its own pull requests (without human supervision) now serve production traffic (model card).

Citadel Securities, Miami, FL — Sep 2025 – Dec 2025

Quantitative Research Intern. Mentor: Dr. Wujie Huang.

- Built GenAI models for alpha signal prediction, improving Pearson IC by more than 5% through learning-algorithm and model architecture design.

Google DeepMind, Mountain View, CA — May 2025 – Aug 2025

Student Researcher. Mentors: Dr. Lun Wang, Dr. Shujian Zhang and Dr. John Lambert.

- Unsupervised discovery of reasoning behaviors in thinking models. Published at ICML 2026 (paper).

Together AI, Remote (part time) — Feb 2025 – May 2025

Research Consultant. Mentors: Dr. Shirley Wu, Dr. Qingyang Wu and Dr. Ben Athiwaratkun.

- Pushed the Pareto frontier of reasoning accuracy versus token cost by steering model activations at test time. Published at NeurIPS 2025 Workshop (paper).

Intel Labs, Remote (part time) Sep 2024 – Jan 2025

Research Intern. Mentors: Dr. Souvik Kundu and Dr. Mostafa Hesham.

- Efficient LLMs with linear attention. Proposed an on-the-fly distillation framework that transforms standard attention into dual-state linear attention, enabling efficient LLM inference. Published at ICML 2025 (paper).
- Calibrating reasoning trajectories of LLMs without additional training, Achieving gains in both efficiency and performance. Published at COLM 2025 (paper).

Meta Reality Labs, Sunnyvale, CA May 2024 – Aug 2024

Research Intern. Mentors: Dr. Yuandong Tian, Dr. Zechun Liu and Dr. Steven Li.

- Efficient LLM inference via activation sparsity, with results published at ICLR 2025. Follow-up implementations contributed to Meta's Ray-Ban smart glasses project (paper).

Microsoft DeepSpeed, Remote (part time) Oct 2023 – Feb 2024

Research Intern. Mentors: Dr. Zhewei Yao and Dr. Shirley Wu.

- Enhanced long-context awareness in LLMs via a training-free head-wise adaptive position encoding strategy. Published at NeurIPS 2024 (paper).

Lawrence Livermore National Laboratory (LLNL), Remote (full time) May 2023 – Aug 2023

Research Intern. Mentors: Dr. Bhavya Kailkhura, Dr. Brian Bartoldson and Dr. James Diffenderfer.

- Efficient long-context LLM inference with KV-cache compression. Published at MLSys 2024 (paper).

EDUCATION

The University of Texas at Austin (UT Austin) Sep 2022 – May 2026

Ph.D. in Electrical and Computer Engineering. Advisor: Prof. Atlas Wang.

Committee members: Prof. Diana Marculescu, Prof. David Z. Pan, Prof. Beidi Chen, Dr. Yuandong Tian.

University of Science and Technology of China (USTC) Sep 2019 – Jun 2022

M.E. in Electrical and Computer Engineering. Advisor: Prof. Bin Li.

University of Science and Technology of China (USTC) Sep 2015 – Jun 2019

B.S. in Applied Physics. Yan Ji-Ci Talent Program in Physics, Minor in Computer Science.

SELECTED PUBLICATIONS

Listed below are several representative publications. The complete list, with over 5.5k citations as of November 2026, is on Google Scholar. A mark * denotes co-first authorship.

Large-scale pre-training and RL: optimization and data

[1] H. Zhu*, **Z. Zhang***, W. Cong, X. Liu, S. Park, V. Chandra, B. Long, D. Pan, Z. Wang, and J. Lee, "APOLLO: SGD-like Memory, AdamW-level Performance", *MLSys*, 2025. **Outstanding Paper Honorable Mention.**
▷ Integrated into HuggingFace, LLaMA-Factory and FluxML.

[2] J. Zhao, **Z. Zhang**, B. Chen, Z. Wang, A. Anandkumar, and Y. Tian, "GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection", *ICML*, 2024. **Oral.**
▷ Integrated into HuggingFace, PyTorch, LLaMA-Factory, Axolotl, FedML and bitsandbytes.

[3] H. Zhu, **Z. Zhang**, H. Huang, D. Su, Z. Liu, J. Zhao, I. Fedorov, H. Pirsiavash, J. Lee, D. Pan, Z. Wang, Y. Tian, and K. Tai, "The Path Not Taken: RLVR Provably Learns Off the Principals", *NeurIPS*, 2026.

[4] S. Xing, J. Hong, Y. Wang, R. Chen, **Z. Zhang**, A. Grama, Z. Tu, and Z. Wang, "LLMs Can Get 'Brain Rot': A Pilot Study on Twitter/X", *COLM*, 2026.
▷ Covered in Nature, WIRED, Forbes and Fortune.

[5] **Z. Zhang***, X. Chen*, T. Chen*, and Z. Wang, "Efficient Lottery Ticket Finding: Less Data is More", *ICML*, 2021.

Efficient inference: caching, sparsity, and long-context modeling

[6] **Z. Zhang**, Y. Sheng, T. Zhou, T. Chen, L. Zheng, R. Cai, Z. Song, Y. Tian, C. Ré, C. Barrett, Z. Wang, and B. Chen, "H_2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models", *NeurIPS*, 2023.
▷ Integrated into Llama-Recipes by Meta; this work inspired Intel and Apple's inference engine.

[7] **Z. Zhang**, Z. Liu, Y. Tian, H. Khaitan, Z. Wang, and S. Li, "R-Sparse: Rank-Aware Activation Sparsity for Efficient LLM Inference", *ICLR*, 2025.
▷ Follow-up implementations contributed to Meta's Ray-Ban smart glasses project.

[8] **Z. Zhang***, S. Liu*, R. Chen, B. Kailkhura, B. Chen, and Z. Wang, "Q-Hitter: A Better Token Oracle for Efficient LLM Inference via Sparse-Quantized KV Cache", *MLSys*, 2024.

[9] **Z. Zhang**, R. Chen, S. Liu, Z. Yao, O. Ruwase, B. Chen, X. Wu, and Z. Wang, "Found in the Middle: How Language Models Use Long Contexts Better via Plug-and-Play Positional Encoding", *NeurIPS*, 2024.

Evaluation and understanding: recursive self-improvement and reasoning behaviors

[10] Y. Zhou, K. Agarwal, Z. Chen, Z. Wang, **Z. Zhang**, W. Li, X. Wei, Z. Zeng, J. Liu, J. Wang, S. Wu, Z. Jiang, Z. Dong, H. Cai, B. Yuan, Y. Sheng, C. Zhang, B. Li, and B. Chen, "Singularity-Bench: Can Agents Build Agents?", *ICLR*, 2027. (*Under review*)

[11] **Z. Zhang**, S. Zhang, J. Lambert, W. Zhou, Z. Wang, M. Chen, A. Hard, R. Mathews, and L. Wang, "Fantastic Reasoning Behaviors and Where to Find Them: Unsupervised Discovery of the Reasoning Process", *ICML*, 2026.

[12] R. Chen*, **Z. Zhang***, J. Hong, S. Kundu, and Z. Wang, "SEAL: Steerable Reasoning Calibration of Large Language Models for Free", *COLM*, 2025.

[13] S. Liu*, T. Chen*, **Z. Zhang**, X. Chen, T. Huang, A. Jaiswal, and Z. Wang, "Sparsity May Cry: Let Us Fail (Current) Sparse Neural Networks Together!", *ICLR*, 2023. **Spotlight.**

Invited talks and academic service

Democratizing LLM Training via Low-Rank Methods — Jul 2025
Intl. Conference on Continuous Optimization (ICCOPT)

Understanding and Steering Cognitive Behaviors in Reasoning Models — Jun 2025
Together AI Turbo Deep Dive Seminar

APOLLO: SGD-like Memory, AdamW-level Performance — Apr 2025
Citadel Securities Ph.D. Summit (invite-only)

Q-Hitter: A Better Token Oracle for Efficient LLM Inference — May 2024
MLSys 2024 Young Professionals Symposium

Efficient Generative Inference by Heavy Hitters and Beyond — Feb 2024
UT Austin Imagine Consortium (MLSys)

H_2O: Heavy-Hitter Oracle for Efficient Generative Inference of LLMs — Jun 2023
Google Sparsity Research Seminar

Conference reviewer: NeurIPS, ICLR, ICML, CVPR, ICCV, ECCV, AISTATS, EMNLP, ACL, ICIP, ICME, CPAL, ACCV, AAAI.

Journal reviewer: TNNLS, JMLR, KBS, Neural Networks, IPM, TCASAI.