Zhaoxiang Feng

Austin, TX | aaronzhfeng@utexas.edu | (336) 337-7139 | aaronzhfeng | aaronzhfeng.github.io

Education

The University of Texas at Austin *Aug. 2026 – Present*
M.S. in Computer Science

University of California San Diego *Sep. 2022 – Mar. 2026*
B.S. in Data Science & B.S. in Probability & Statistics

Publications

[1] Matthew Ho, Chen Si, **Zhaoxiang Feng**, Fangxu Yu, Yichi Yang, Zhijian Liu, Zhiting Hu, Lianhui Qin. "ArcMemo: Abstract Reasoning Composition with Lifelong LLM Memory."
***Runner Up**, ARC Prize 2025 Paper Awards, Top 8 of 90 submissions.* *[paper] [code]*

[2] Lanxiang Hu, **Zhaoxiang Feng**, Yulun Wu, Haoran Yuan, Yujie Zhao, Yu-Yang Qian, Bojun Wang, Peng Zhao, Daxin Jiang, Yibo Zhu, Tajana Rosing, Hao Zhang. "JetSpec: Breaking the Scaling Ceiling of Speculative Decoding with Parallel Tree Drafting."
Under review at the Conference on Neural Information Processing Systems (NeurIPS), 2026. *[paper] [code]*

Research Experience

Research Assistant | Hao AI Lab, UC San Diego *Sep. 2025 – Present*
Advisors: Prof. Hao Zhang and Lanxiang Hu

- Led **JetSpec++**, a causal correction for parallel drafting that conditions each drafted token on its full branch history rather than on the preceding token alone; at matched parameter count it outperformed DeepSeek's DSpark, with the margin widening at draft depths beyond those used in training.
- Contributed to **JetSpec**, a draft head that generates a complete speculative tree in one forward pass while preserving the target model's autoregressive factorization along every branch; built its standalone inference engine (paged KV cache, Triton tree-attention kernel, CUDA-graph verification) and its evaluation suite. JetSpec reaches up to 9.64× end-to-end speedup over autoregressive decoding on Qwen3-8B (MATH-500).
- Ported **DFlash** speculative decoding to TPU in JAX as the primary contributor, showing that verification cost stays flat in draft block size through K=1024. Averaged 3.13× speedup on TPU v5p; the draft model and proposer are merged into vLLM's TPU backend with pipeline integration in review, and the work is featured on the Google Developers Blog.
- Built the training stack for a 1.8B-parameter language model on eight NVIDIA B200 GPUs, implementing the full pre-training and SFT and DPO post-training system with DDP and ZeRO-1.

Research Assistant | Q-Lab, UC San Diego *Jan. 2025 – Jun. 2026*
Advisors: Prof. Lianhui Qin and Matthew Ho

- Contributed to the design of **ArcMemo**'s program-synthesis memory ontology, a typed, parameterized representation in which reusable reasoning concepts are stored and composed.
- Worked on the reasoning-guided concept selection that replaced embedding-based retrieval after embedding search proved ineffective on the ARC-AGI-1 abstract-reasoning benchmark.
- Built the pipeline that produces concept-labeled helper puzzles: each hand-written concept is expanded by an LLM into puzzle-generating code, which is executed and tested so that every generated puzzle is verified to exercise its concept.
- Designed the framework's adaptation to AIME competition mathematics and ran the generalization experiments and memory-format comparisons that informed the final design.

Technical Skills

Languages: Python, R, SQL, Java, JavaScript

ML frameworks: PyTorch, JAX, Hugging Face Transformers

Systems & infrastructure: LLM serving engines (vLLM; JetSpec's JetFlow engine), distributed GPU training, TPU inference (JAX), Docker