Yufa Zhou

☎ +1 445 260 7203 | ✉ yufa.zhou@duke.edu | Homepage | GitHub | Google Scholar

Education

Duke University, Durham, United States — Aug. 2025 — May. 2030 (expected)
Ph.D. Student in Computer Science
Advisor: **Anru Zhang**

University of Pennsylvania, Philadelphia, United States — Aug. 2023 — May. 2025
M.S.E. in Scientific Computing — GPA: 3.97/4.00

Wuhan University, Wuhan, China — Sep. 2019 — Jul. 2023
B.E. in Engineering Mechanics — GPA: 3.68/4.00

Skills

AI Expertise: LLMs, Generative AI, Multi-Agent Systems, Deep Learning Theory, Physics of LLMs
Technical Skills: Prompt Engineering, Post-Training (SFT, RLHF, RLVR), Inference & Model Optimization, AI Interpretability, Theoretical & Empirical Analysis, Statistics
Programming: Python (PyTorch, JAX, HuggingFace), Linux, SQL, Git, LaTeX

Research Experiences

Mentors: **Shuyan Zhou**, **Anru Zhang** — Duke, Sep. 2025 — Present

- Conducting **foundational research** in **interpretability**, analyzing **LLM reasoning** through a **differential-geometric framework of reasoning flows** that models embeddings as smooth trajectories on concept manifolds, where logic emerges as differential constraints on velocity and curvature. Demonstrating consistent geometric invariants across topics, languages, and model families, offering a unified lens into how reasoning unfolds within large models. [Paper] [GitHub] [Submitted to ICLR 2026]

Mentors: **Surbhi Goel**, **Anru Zhang** — UPenn, Duke, Dec. 2024 — Present

- Conducting theoretical research on **ML/DL theory**, establishing the first **fundamental limits of Transformers** for time-series forecasting under AR(p) processes through in-context learning theory, supported by controlled synthetic experiments and ablation studies. Accepted as an **Oral (3/68, ≈4.4%)** at the **WCTD Workshop, NeurIPS 2025**. [Paper] [GitHub] [Submitted to ICLR 2026]

Mentors: **Xuan Shen**, **Yanzhi Wang**, **Jiuxiang Gu** — Remote, Aug. 2024 — Present

- Researching **efficiency and acceleration of Transformers and generative models**, including: a training-free numerical pruning method using Newton-based importance scoring with compensation for autoregressive model compression (**AAAI 2025** [Paper]), LazyDiT for computation reuse and redundancy skipping in diffusion transformers (**AAAI 2025** [Paper]), FastCar for cache-attentive replay in autoregressive video generation on edge devices ([Paper]), and DraftAttention for low-resolution–guided sparsity in video diffusion transformers ([Paper]).

Projects

MASSE: Building Multi-Agent Systems for Real-World Workflow Automation — May. 2025 — Oct. 2025

- **Lead author.** Co-conceived and implemented the first LLM-driven multi-agent system (**MASSE**) that automates end-to-end structural design workflows. Designed the three-team architecture (*Analyst–Engineer–Manager*) using **AutoGen** for orchestration, structured JSON I/O, and persistent agent memory; integrated FEM solvers, engineering codes, and tool-augmented reasoning for transparent, verifiable analysis. Released open-source code [GitHub] and preprint [Paper] .

- **Experiments & evaluation.** Designed and conducted benchmarking on domain-grounded datasets, demonstrating a 98% reduction in expert workload (132 → 2 min) and consistent performance across GPT-4o, and o4-mini backbones. Analyzed cost–latency trade-offs and performed ablations on structured memory and I/O components to validate system scalability and reasoning reliability.

Recon: Post-Training LLMs for Economic & Multi-Agent Reasoning March. 2025 — May. 2025

- **Lead author.** Proposed that domain-aligned post-training (**SFT** + **GRPO** under **RLVR**) can induce *strategic generalization* without interactive data. Built the full pipeline with reasoning-trace distillation, hierarchical reward design, and a curated 2,100-problem, 15-category dataset; coordinated cross-institution collaboration; released [GitHub] & [Paper].
- **Experiments.** Fine-tuned the DeepSeek-R1-Distill-Qwen-7B model using the **Unsloth** library on an **NVIDIA H100 GPU**, employing LoRA-based SFT and GRPO via **TRL**. Benchmarked on economic reasoning and multi-agent games (self-play & adversarial), achieving +14.7 pp accuracy and +9.5 pp Nash-equilibrium convergence. Performed SFT vs. RL ablations and analyzed emergent equilibrium-seeking behavior, showing post-training as a scalable path to structured reasoning and agent alignment.

Selected Publications

(Indicates alphabetical order or equal contribution)*

Accepted in the peer-reviewed venues:

1. Zichen Wen, Shaobo Wang, **Yufa Zhou**, Junyuan Zhang, Qintong Zhang, Yifeng Gao, Zhaorun Chen, Bin Wang, Weijia Li, Conghui He, Linfeng Zhang. "Efficient Multi-modal Large Language Models via Progressive Consistency Distillation." In *NeurIPS*, 2025. [Paper]
2. Yingyu Liang*, Jiangxuan Long*, Zhenmei Shi*, Zhao Song*, **Yufa Zhou***. "Beyond Linear Approximations: A Novel Pruning Approach for Attention Matrix." In *ICLR*, 2025. [Paper]
3. Yingyu Liang*, Zhizhou Sha*, Zhenmei Shi*, Zhao Song*, Mingda Wan*, **Yufa Zhou***. "Unraveling the Smoothness Properties of Diffusion Models: A Gaussian Mixture Perspective." In *ICCV*, 2025. [Paper]
4. Xuan Shen, Zhao Song, **Yufa Zhou**, Bo Chen, Yanyu Li, Yifan Gong, Kai Zhang, Hao Tan, Jason Kuen, Henghui Ding, Zhihao Shu, Wei Niu, Pu Zhao, Yanzhi Wang, Jiuxiang Gu. "LazyDiT: Lazy Learning for the Acceleration of Diffusion Transformers." In *AAAI*, 2025. [Paper]
5. Xuan Shen, Zhao Song, **Yufa Zhou**, Bo Chen, Jing Liu, Ruiyi Zhang, Ryan A. Rossi, Hao Tan, Tong Yu, Xiang Chen, Yufan Zhou, Tong Sun, Pu Zhao, Yanzhi Wang, Jiuxiang Gu. "Numerical Pruning for Efficient Autoregressive Models." In *AAAI*, 2025. [Paper]
6. Yingyu Liang*, Zhizhou Sha*, Zhenmei Shi*, Zhao Song*, **Yufa Zhou***. "Looped ReLU MLPs May Be All You Need as Practical Programmable Computers." In *AISTATS*, 2025. [Paper]

Preprints:

1. **Yufa Zhou***, Yixiao Wang*, Xunjian Yin*, Shuyan Zhou, Anru R. Zhang. "The Geometry of Reasoning: Flowing Logics in Representation Space." arXiv preprint arXiv:2510.09782, 2025. [Paper] [Submitted to ICLR 2026]
2. **Yufa Zhou***, Yixiao Wang*, Surbhi Goel, Anru R. Zhang. "Why Do Transformers Fail to Forecast Time Series In-Context?" Oral (3/68, ≈4.4%) at WCTD Workshop, NeurIPS 2025. [Paper] [Submitted to ICLR 2026]

Academic Services

Conference Reviewer: ICLR (2025, 2026), NAACL 2025, ACL 2025, EMNLP 2025, AAAI 2026.
Journal Reviewer: TKDE, TNNLS.