Zhaoxiang (Aaron) Feng

San Diego, CA | zhf004@ucsd.edu | (336) 337-7139 | aaronzhfeng | aaronzhfeng.github.io

Education

University of California San Diego (UCSD) *Sep. 2022 – Jun. 2026*
B.S. in Data Science & B.S. in Probability & Statistics *(expected)*

- GPA: 3.83 (Major GPAs: 3.95 in Data Science, 3.93 in Statistics)
- Honors: Provost Honors (multiple quarters)
- Graduate-level Coursework: Advanced Time Series Analysis, Machine Learning, Applied Statistics, Privacy-sensitive Data Systems

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."
Under review at International Conference on Learning Representations (ICLR), 2026. *[paper] [code]*
***Runner Up**, ARC Prize 2025 Paper Awards — Top 8 of 90 paper submissions.*

[2] **Zhaoxiang Feng**, David Scott Lewis, Enrique Zueco. "LabMemo: Concept-Level Memory for Autonomous Scientific Discovery."
Accepted at IEEE IROS 2025 Workshop on Embodied AI and Robotics for Future Scientific Discovery (AIR4S) (non-archival), 2025. *[paper] [code]*

[3] **Zhaoxiang Feng**, David Scott Lewis. "SOKRATES: Distilling Symbolic Knowledge into Option-Level Reasoning via Solver-Guided Preference Optimization."
Accepted at AAAI 2026 Bridge Program on Logic & AI: Logical and Symbolic Reasoning in Language Models (LMReasoning), 2026. *[paper] [code]*

Research Experience

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

- Contributed to ArcMemo by leading the design of a program-synthesis-style memory ontology, developing many-to-one puzzle-to-feature mappings and manually curating concept parameterizations to enable concept-level reasoning.
- Engineered a reasoning-based retrieval mechanism (System-2 exploration) to resolve embedding failures, achieving a 7.5% relative gain on ARC-AGI-1 (59.33% official score).
- Built a complete concept dataset generation pipeline transforming hand-written concepts into validated helper puzzles through multi-stage LLM-based generation, code synthesis, and automated testing.
- Extended the framework to AIME math problems by designing a metacognitive self-assessment pipeline, improving accuracy by 9.3% via self-reflective memory usage.

Research Assistant | Wang Lab, UC San Diego *Jun. 2025 – Present*
Advisors: Prof. Wei Wang and Young Su Ko

- Developing ProteinAgent (Ongoing): An agentic framework orchestrating domain tools (AlphaFold, MD simulations) using LabMemo-style concepts to guide protocol design for protein stability analysis.
- Architected a heterogeneous Mixture-of-Experts system for chemical reaction prediction, integrating four specialized expert models (graph-based, SMILES sequence, 3D geometry-aware, and condition-aware) with learned routing mechanisms on USPTO reaction datasets.
- Implemented graph neural network encoders using directed message-passing architectures with shortest-path positional encodings, enabling permutation-invariant molecular representations for stereochemistry-sensitive reaction modeling.
- Engineered training pipelines with teacher-forcing, load-balancing losses, and router warmup; designed evaluation frameworks for top-k accuracy and per-expert ablation.

Student Researcher | Halıcıoğlu Data Science Institute, UC San Diego *Sep. 2025 – Present*
Advisor: Prof. Hao Zhang

- Senior Capstone: "Open LLM Training, Inference, and Infrastructure." Building scalable pipelines for training open-source large language models, benchmarking inference frameworks, and exploring alignment strategies.

Research Assistant | U.S. Immigration Policy Center, UC San Diego *Sep. 2023 – Jun. 2024*
Advisors: Prof. Tom K. Wong and Dr. Gabriel De Roche

- Utilized Python and R to analyze historical naturalization data and forecast long-term immigration trends, informing policy briefs for the 2024 election cycle.
- Developed data pipelines and visual dashboards to communicate findings to non-technical stakeholders.

Teaching Experience

Tutor — DSC 40A: Theoretical Foundations of Data Science I *Summer 2025*
University of California San Diego

- Conducted tutoring sessions covering set theory, probability, and algorithmic thinking for early-career data science students.
- Led review sessions and prepared supplementary materials to aid student learning.
- Supervised Oral Exam

Grader — Mathematics Department *Fall 2024 – Fall 2025*
University of California San Diego

- Graded for MATH 180A/C (Probability & Stochastic Processes), MATH 181A (Mathematical Statistics), and MATH 185 (Computational Statistics) a total of 7 times.

Technical Skills

Languages: Python, Java, R, SQL, JavaScript, HTML

ML Frameworks: PyTorch, PyTorch Geometric, scikit-learn, Transformers, NVIDIA PhysicsNeMo, neural operators (Fourier Neural Operator)

LLM/NLP: OpenAI APIs, prompt engineering, LLM fine-tuning (PPO, DPO), RAG systems, memory-augmented LLMs, reasoning-based retrieval, vision-language models

Data & Infrastructure: Pandas, Dask, Spark, AWS, Docker, distributed training, vector databases (embedding retrieval)

Dev Tools: Git/GitHub, Linux, Hydra, SSH

Scientific ML & Agents: physics-based ML (PDE/inverse problems), continual learning (Elastic Weight Consolidation), concept-level memory design, autonomous scientific agents, safety-aware Planner/Selector/Verifier architectures