Mrinaal Arora

Mumbai, India

aroramrinaal@gmail.com • +91 99677 44102 • aroramrinaal.com • linkedin.com/in/aroramri

Education

Arizona State University **Aug 2022 – Aug 2025**

Bachelor of Science in Data Science, summa cum laude; GPA: 3.84/4.00 *Tempe, Arizona*

Selected Independent Research Projects

(8) **Emergent Misalignment in a 3.35B Language Model** [write-up] [code] [artifacts] **August 2026**

AI alignment, parameter-efficient fine-tuning, controlled evaluation, and behavioral safety

- Tested whether narrow risky-financial fine-tuning produces broad emergent misalignment in Tiny Aya Global by training three independent rsLoRA adapters on 6,000 examples and a matched prudent control with identical prompts, row order, model, and training recipe.
- Generated and blindly judged 4,320 paired responses per run across 72 open-ended prompts; risky adapters averaged 19.37% non-financial emergent misalignment versus 2.03% for base models, a +17.33 percentage-point shift across three seeds, while the prudent control changed by only +0.67 points.

(7) **World Fan Arena: A Multi-Agent Football Fan Simulation** [write-up] [live] **July 2026**

Multi-agent systems, persistent simulations, and event-driven agent infrastructure

- Built a public simulation in which AI fan agents representing different countries discuss live football matches, rivalries, predictions, upsets, and evolving tournament narratives in persistent chat rooms.
- Designed an event-driven pipeline using heterogeneous language models, live match data, durable state, queues, and scheduled updates every five minutes to coordinate long-running agent interactions.

(6) **Efficient Qwen: Low-Latency Qwen3.5-4B Inference on a Single A10G** [write-up] **June 2026**

Efficient inference, quantization, speculative decoding, and model-serving systems

- Conducted more than 100 controlled experiments to reduce Qwen3.5-4B inference latency on a single NVIDIA A10G for the ICML 2026 AdaptFM competition.
- Combined GPTQ INT4 quantization, speculative decoding, prefix caching, and serving optimizations to achieve a 4.342× speedup over the competition baseline.

(5) **Looking Inside Qwen3-4B with Sparse Autoencoders** [write-up] [code] **June 2026**

Mechanistic interpretability, sparse feature discovery, and causal steering

- Trained a 10,240-feature sparse autoencoder on layer-20 activations from 50M FineWeb-Edu tokens, compared multiple sparsity settings, and automatically labeled 9,418 learned features.
- Identified feature 3311 as associated with cooking-instruction language and tested its causal influence using activation steering; moderate steering shifted neutral generations toward procedural cooking language, while stronger steering produced repetitive, degenerate outputs.

(4) **Pretraining a 124M Language Model from Scratch** [write-up] [models] **May 2026**

Language-model architecture, pretraining, continued pretraining, and evaluation

- Implemented a 123.6M-parameter decoder-only transformer with RoPE, RMSNorm, SwiGLU, and tied embeddings, then trained it from random initialization on 2B FineWeb-Edu tokens using a single H100.
- Achieved a minimum validation loss of 3.45 and published both checkpoints. A subsequent 1B-token continued-pretraining run did not improve HellaSwag accuracy; both checkpoints scored 28.44%.

(3) **Parameter Golf: Language-Model Training Under 16 MB** [write-up] [code] **May 2026**

Parameter-efficient pretraining, constrained optimization, and controlled experimentation

- Ran controlled H100 experiments for OpenAI Model Craft under a 16 MB compressed-artifact limit and a ten-minute training budget, testing context lengths, architectural changes, recurrence, and evaluation strategies.
- Improved the single-H100 baseline from 1.36785 to 1.28397 bits per byte while keeping the final compressed artifact within the competition size constraint.

Selected Independent Research Projects — Continued

(2) **Learning Wordle with GRPO** [write-up] [model] **April 2026**

Reinforcement learning from verifiable rewards and environment-based evaluation

- Fine-tuned Qwen3-1.7B with GRPO for 62 training steps using four rollouts per prompt and approximately 15,800 Wordle games, with shaped rewards for valid guesses, letter placement, and non-repetition.
- Evaluated each model over 50 episodes; the trained model won 4/50 episodes (8%), compared with 0/50 for the base model, though the sample was small.

(1) **Cold-Start Reasoning SFT for Nanbeige4-3B** [write-up] [code] [models] **March 2026**

Reasoning-trace distillation, parameter-efficient fine-tuning, and failure analysis

- Fine-tuned Nanbeige4-3B-Base with LoRA on an H100 using 2,160 distilled reasoning triplets, then scaled the experiment to 12,000 examples over two epochs and published the resulting adapters.
- Analyzed reasoning traces from multiple source models and found that 78.8% of examples in the larger dataset were truncated at 2,048 tokens, likely explaining the trained model's frequent failure to complete its reasoning trace or produce a final answer.

Professional Experience

Ohara [Work] **May 2025 – Jan 2026**

Software Engineer *Remote*

- Led frontend development for Ohara's main product and shipped it as a live **Base Mini App** on join.base.app.
- Built the full app-embedding pipeline for AI-generated applications, enabling them to run natively within the Base ecosystem.
- Implemented user authentication using **Sign in with Farcaster**, Quick Auth, and the Farcaster Mini App SDK across both web and embedded surfaces.

Ohara **Jan 2025 – May 2025**

Software Engineering Intern *Remote*

- Implemented production integrations with OpenAI, OnchainKit, Helius, and World App to power AI-based app generation.
- Built and shipped full-stack features such as the App Detail Page and Frames integration, allowing Ohara apps to be embedded on the Warpcast platform.
- Contributed more than 5,000 lines of production code across 15+ pull requests on a remote, asynchronous engineering team.

Biodesign Center for Applied Structural Discovery, Arizona State University **Jan 2023 – Nov 2023**

Research Lab Aide *Tempe, Arizona*

- Supported structural biology research on protein structure prediction and photosynthesis, including cryo-EM data processing and analysis using RELION.
- Performed SDS-PAGE, BCA assays, Western blotting, and Coomassie staining while following laboratory safety protocols.
- Prepared stock solutions, managed chemical inventory and laboratory supplies, and assisted researchers with day-to-day laboratory operations.

Selected Achievements and Awards

- Created iammusic2024.com, reaching 157K+ lifetime unique visitors [write-up].
- Built and shipped web experiments that collectively reached 200K+ unique visitors in a 12-month period, including traffic to iammusic2024.com [write-up].
- Received the New American University merit scholarship upon admission to Arizona State University and was named to the Dean's List four times during undergraduate study.
- Won five hackathons during undergraduate study, including YHacks at Yale (Sustainability Track Winner) and ASU events Hack SoDA 2024 (Gradify) and DevHacks (RenVue). Also placed Runner-Up at MakeHarvard 2024 for PowerStride [write-up].

Technical Skills

AI Tools: Claude Code, Codex, Cursor Agent, Vercel v0
ML and Research: PyTorch, Hugging Face Transformers, TRL, PEFT, vLLM, Hugging Face Datasets
Databases: PostgreSQL
Web and Infrastructure: Docker, Next.js, Prisma, React, Tailwind CSS, Neon, Railway, Vercel
Languages: C++, Go, Python, TypeScript