

Colin Zhao

+1 (914)-223-0536 | colinz@seas.upenn.edu | Website | LinkedIn | Github

EDUCATION

University of Pennsylvania

Bachelor of Engineering, Major in Computer Science

Expected May 2028

GPA: 3.6

- Coursework: Applied ML*, Artificial Intelligence*, Databases*, Big Data Analytics*, Algorithms, Linear Algebra (*Graduate Course)

TECHNICAL SKILLS

Languages: Python, JavaScript/TypeScript, C, C++, HTML/CSS, Java

Tools & Frameworks: PyTorch, AWS, Linux, React, Next.js, PostgreSQL, Docker, Playwright, HuggingFace

EXPERIENCE

Member of Technical Staff Intern

Refresh (YC X25)

February 2026 - Present

San Francisco, CA

- Built the **Gemini** computer-use support and deterministic reward verifier for an internal computer-use agent, later open-sourced as part of the **Harbor** framework; supports live-web and archived WACZ execution
- Created pipeline to automate cloning an EHR software by automatically scraping training videos and documents into structured UI specs and executable tests and delegating agents to build the clone those artifacts enforce
- Designed and shipped 40 open-web **RLVR** tasks with deterministic reward signals for training and evaluating web agents

Researcher

Penn Distributed Systems Laboratory

January 2026 - May 2026

Philadelphia, PA

- Developed 2 full-stack benchmarking websites and 200 RLVR tasks with deterministic backend checks targeting time-sensitive features to assess web agent robustness; paper under review at NeurIPS 2026

Researcher

Penn Medicine

October 2025 - May 2026

Philadelphia, PA

- Built an open-source multi-agent framework with self-improvement, achieving **state-of-the-art 66.2%** on the PubMedQA benchmark (+**4.6 points** over the strongest baseline); accepted as a poster at **RSNA 2026**
- Finetuned Qwen3-4B on synthetic chest CT radiology reports using **SFT + GRPO** reinforcement learning, achieving **47%** improvement in impression quality over baseline across 17 thoracic pathologies

Software/Data Center Intern

Antalpa

June 2025 - August 2025

Atlanta, GA

- Developed **terminal agents** to autonomously configure 630+ water-cooled GPU clusters via NVIDIA SMI, reducing manual setup time by **75%**

STARTUPS/PROJECTS

WorkFleet | *FastAPI, React, TypeScript*

- Built a control-plane webapp orchestrating a fleet of cloud VMs, each running native desktop apps driven by a computer-use agent, enabling concurrent GUI-automation tasks on isolated remote desktops via FIFO dispatch and WebSocket-streamed live screenshots
- Designed an “app-state mirror” syncing a reference machine’s logged-in profiles to disposable workers through daily, **SHA-256**-verified versioned snapshots with upload-then-flip publishing and atomic restore

Collegize.ai | *Typescript, React, Next.js, AWS*

- Built and scaled SaaS platform to **\$100k+ total sales** and 500+ paying users
- Engineered full-stack backend infrastructure with 100+ **Next.js** API routes integrating Stripe payments, Gemini AI, **AWS S3** document storage, and **PostgreSQL** via Supabase

Foil | *Python*

- Built an AI text detection model using DeBERTa-v3-base with iterative hard negative mining, scanning for false positives, reverse-engineering prompts via Gemini, and generating targeted training mirrors
- Outperforms GPTZero on the checkforai benchmark (95.2% accuracy vs 94.2%)

ScoundReL — Card Game RL Environment & Agent | *Python, Gymnasium*

- Implemented a game engine and Gymnasium environment, trained Maskable PPO (sb3-contrib) with (~50M env steps, 8 parallel envs) with **vectorized** rollouts, linear LR schedule, masked eval, achieving a **75%** win rate