Colin Zhao

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

Education

University of Pennsylvania — Philadelphia, PA
Bachelor of Engineering, Major in Computer Science, Minor in Economics — *Expected December 2027*

- Relevant Coursework: Database and Information Systems (Graduate), Big Data Analytics (Graduate), Data Structures and Algorithms, Econometric Methods, Linear Algebra
- 3.6/4.0 GPA

Technical Skills

Languages: Python, JavaScript/TypeScript, HTML/CSS, Java, SQL
Tools & Frameworks: PyTorch, AWS, React, LangChain, PostgreSQL, Git, Node.js, ROS2, OpenCV, Linux

Experience

Researcher — December 2025 - Present
Penn Distributed Systems Laboratory — *Philadelphia, PA*

- Implemented Quest KV-cache optimization in SGLang-based serving engine, reducing long-context LLM inference latency by **30%** for 128k+ token workloads
- Developed benchmarking infrastructure for web agents, creating a Yahoo Finance-style test website to evaluate agent performance on dynamic UI patterns including hover states and scroll-based data loading
- Contributing to research publication on agentic AI browser optimization techniques, developing custom evaluation framework and analyzing Quest KV-cache performance across 15+ long-context benchmarks

Researcher — October 2025 - Present
Penn Medicine — *Philadelphia, PA*

- Engineered multi-agent AI orchestration system using **MedGemma** for radiology report generation, implementing staged reasoning chains with integrated QA validation to achieve clinical-grade accuracy on 500+ test cases
- Evaluated system on MedHELM, achieving a 15% increase in accuracy compared to the base Gemini model
- Developed real-time clinical documentation pipeline converting live physician dictation into structured OpenEHR records and radiology reports using MedASR and models optimized for low-resource clinical settings

Software Developer — September 2025 - December 2025
Penn Aerial Robotics — *Philadelphia, PA*

- Engineered real-time UAV computer vision system integrating OpenCV payload detection (30 FPS, 95% accuracy) with ROS2-based position estimation from camera data, enabling autonomous navigation for competition drones

Software/Data Center Intern — June 2025 - August 2025
Antalpha — *Atlanta, GA*

- Architected autonomous deployment system for 630+ water-cooled GPU clusters using LangChain agents and NVIDIA SMI, reducing manual configuration time by **75%**

Startups/Projects

Founding Engineer at Collegize.ai | *Typescript, React, Next.js, AWS*

- Scaled SaaS platform to **$8k+ MRR** with **350+ paying customers**, achieving 40% month-over-month growth through product-led acquisition
- Engineered full-stack backend infrastructure with 50+ **Next.js** API routes integrating Stripe payments, Gemini AI, **AWS** S3 document storage, and **PostgreSQL** via Supabase, processing 1,000+ application documents monthly
- Shipped 2 MVPs and managed entire product development lifecycle, from ideation to deployment with **Vercel**

Maurice Macros - Nutrition Tracking Mobile App | *SQL, React Native*

- Built full-stack mobile nutrition tracking app using React Native, Node.js, and PostgreSQL with USDA API integration, Clerk authentication, and RESTful backend

Real-Time Screen Gaze Tracker | *Python*

- Developed real-time gaze tracking system using MediaPipe facial landmarks and custom geometric transformations, achieving 95% accuracy within 3° visual angle for UX research and accessibility applications
- Implemented session recording functionality for eye-tracking research applications and user attention analysis

DBKeeper - Desktop App for SQLite File Interaction | *Javascript, SQL*

- Built cross-platform desktop application for SQLite database management with query execution and CSV export using Electron.js