KEVIN ZHU

New York, NY | (657) 331-0806 | kevin@kevin-zhu.com | linkedin.com/in/kevinjzhu

EXPERIENCE

SCHONFELD **New York, NY**

Senior Software Engineer – Applied AI Solutions *Jan 2025 – Present*

- Drive expansion of Schonfeld's AI platform, evolving it from a basic chatbot into an expansive AI ecosystem **used daily by >70% of the firm**—managing stakeholders and coordinating cross-functional teams to ensure success
- Designed and launched a custom chatbot framework with **Retrieval-Augmented Generation (RAG)** using LiteLLM, OpenAI, Bedrock, and Kendra—enabling teams to build agentic AI assistants in a **fully self-service model**
- **Lead company-wide AI enablement**, presenting training sessions to executives, traders, analysts, quants, and research teams, driving adoption and integration of AI across business functions at all levels
- Act as an **internal AI expert**, advising teams on model selection, retrieval strategies, AI safety, and emerging techniques, accelerating adoption and guiding AI strategy
- Engage with users to identify pain points and translate feedback into **intuitive UX improvements**—such as model customization, saved prompts, and extensive file support—making AI tools more seamless and approachable

Software Engineer – Applied AI Solutions *June 2023 – Jan 2025*

- Developed internal **LLM-powered AI assistant**, integrating with proprietary data sources, automation tools, and internal APIs to streamline workflows
- Partnered with stakeholders to develop **AI-driven automation solutions** that enhanced quant recruiting, risk reporting, legal document review, market research, etc.
- Pioneered **agentic workflows** at Schonfeld by developing a function-calling AI assistant for IT support—serving as the blueprint for firm-wide AI expansion
- Implemented **real-time AI performance monitoring and testing** with Datadog, incorporating latency tracking, sentiment analysis, topic clustering, and failover mechanisms

AMAZON WEB SERVICES (AWS) **Seattle, WA**

Software Development Engineer I – Elastic Compute Cloud (EC2) *Mar 2022 – June 2023*

- Architected and deployed major upgrade to the **EC2 ReportInstanceStatus API**, adding automated remediation, data analytics, and Grafana dashboarding while optimizing for performance and scalability
- Led successful cross-functional effort reducing EC2 server impairment recidivism from 7% to 2%, resulting in infrastructure efficiency gains of **>$10 million annually**
- Diagnosed and fixed long-standing issue causing false EC2 maintenance alerts on 42,000 instances / month, preventing unnecessary customer pain and earning recognition for **customer obsession**

Software Development Engineer Intern – Elastic Compute Cloud (EC2) *May 2021 – Aug 2021*

- Built server health analytics pipeline covering **>4 million EC2 instances** and a dashboard to visualize failure signals, reducing new signal onboarding time by >80%
- Automated ticket creation for repeat EC2 failures, reducing manual toil and preventing bad hosts from serving customers

BILL **Palo Alto, CA**

Software Engineering Intern – Data & Partner Integrations *June 2020 – Jan 2021*

- Developed and launched new Angular status page, enhancing error visibility and self-service troubleshooting for customers, used by >70% of customers
- Designed and implemented backend Java endpoints, enabling fast and flexible filtering and sorting of customer data

EDUCATION

UNIVERSITY OF CALIFORNIA, BERKELEY **Berkeley, CA**

B.A. in Computer Science (GPA: 3.6/4.0) *Class of 2021*

SKILLS

Languages: Python, Java, SQL, TypeScript, HTML/CSS
Frameworks & Tools: OpenAI, Anthropic, Bedrock, MCP, LangChain, LiteLLM, FastAPI, Flask, Datadog, GitHub Actions
Cloud & Infrastructure: AWS (EC2, S3, DynamoDB, Bedrock, Lambda, Kinesis, Redshift), Terraform, Kubernetes
Hobbies & Interests: Film photography, mixology, downhill skiing, grilling meat