Weici (Will) Zhao

Philadelphia, PA | +1 (434) 257-4183 | williamzhaowc@gmail.com

Software Engineer | Full-Stack Development, AI Agents & RAG

EDUCATION

University of Pennsylvania — *Philadelphia, PA*
M.S. in AI for Data Science | GPA: 3.78 / 4.0 — Expected May 2027

University of Virginia — *Charlottesville, VA*
Dual Bachelor's Degrees: Computer Science & Applied Statistics | GPA: 3.91 / 4.0 — Aug. 2021 – Dec. 2024

PROFESSIONAL EXPERIENCE

APPLE — *Beijing, China | Mar. 2026 – Sep. 2026*
AI Engineer Intern, Sales (via third-party agency)

- Implemented a **human-in-the-loop review agent** that automated approximately **95% of submission reviews**, with no false approvals observed across **90+ validation cases**.
- Built **frontend workflows and backend integrations** for a retail experimentation platform with A/B-test assignment, submission intake, monitoring, and automated reminders; queried dealer and store data with **SQL**.
- Developed **JD.com competitive-audit automation** using XCUITest, OCR, and vision LLMs, covering **12 accessory categories and 25✦ product-search scenarios** with fallback paths for UI and search failures; built a supporting web dashboard and extended audit coverage to U.S. Amazon through a scraping workflow.
- Built a **local-first spreadsheet agent** to split and merge semi-structured Excel workbooks using schema inference and **validated execution plans**; added an evaluation harness, plan caching, and an API for agent-to-agent orchestration.

DICA — Applied AI Startup — *Shanghai, China | Mar. 2025 – Jul. 2025*
AI Engineer Intern

- Built a tool-using **ReAct agent for tabular data analysis** with plot generation and natural-language interpretation, improving accuracy by 7 percentage points over the o4-mini baseline on DABstep's Easy split.
- Identified **stale-memory leakage** across context boundaries in role-playing LLM agents and developed **context masking** to limit reuse of information outside the intended context.
- Built a role-playing LLM **evaluation pipeline** with persona extraction, question templates, and cross-model voting; curated a **100-persona Pantheon-derived dataset** to compare self-consistency and contradiction rates.

SUPREMIND — *Hangzhou, China | May 2024 – Jul. 2024*
Software Engineer Intern — CAD Automation

- Developed a **C# automation tool using Siemens NX Open APIs** and a parameter-entry interface, consolidating customer modeling workflows into repeatable sequences that **generate complete models with one click**.

PERSONAL PROJECTS

WeChat2All - Extensible Agent Control Plane

- Built an **open-source, local-first agent integration platform** using **TypeScript, Node.js, React, and Tauri**, connecting mobile messaging to GUI agents, CLIs, backend APIs, and MCP tools.
- Designed a **plugin protocol and SDK** for isolated, hot-installable integrations, allowing tools and workflows to be added or replaced without modifying or redeploying the core platform.

Legal Manus - RAG-Powered Legal Q&A Platform | Website

- Built and deployed a **RAG-based legal Q&A application** grounded in official documents, with **FastAPI streaming APIs** and a **React frontend** for interactive legal research.
- Indexed **600K+** legal passages using OpenAI text-embedding-3-small and added a BGE cross-encoder reranker, increasing **Hit Rate@3 from 68% to 98%** on a 200-question evaluation set.

TECHNICAL SKILLS

- ***Languages:*** Python, TypeScript, SQL, Java, C++, C#
- ***Full-Stack & Data:*** React, Node.js, FastAPI, Spring Boot, REST APIs, PostgreSQL, MySQL, Redis, Supabase
- ***AI & ML:*** PyTorch, Transformers, LangChain, RAG, Tool Calling, Embeddings, Vector Search, Reranking, LLM Evaluation, Vision LLMs, OCR, MCP, scikit-learn, pandas, NumPy
- **AI Coding Tools:** Codex, Claude Code, Cursor, GitHub Copilot (planning, development, testing, code review)
- ***Cloud & DevOps:*** Docker, AWS EC2, Alibaba Cloud OSS/RDS, Git, CI/CD, Linux