

Yiwen Ding

Philadelphia, PA | dingywn@engineering.upenn.edu | 607-379-1110
LinkedIn: dingyiwen | Portfolio | Graduation: May 2027
Work Authorization: STEM OPT | Active H-1B visa (5+ yrs remaining)

Technical Skills

- **Software & Systems:** **Java, C++, Python**; TypeScript, JavaScript, HTML/CSS, Node.js, React/Next.js, React Native, FastAPI, REST APIs, pytest/JUnit; SQL (SQLite, PostgreSQL), Supabase; Redis/BullMQ, Cloudflare Workers, Akamai; WebAssembly, Linux/Bash, OAuth/SSO; Distributed Systems, Concurrency, RPC, Fault Tolerance, System Design; Git, Docker, CI/CD, AWS.
- **AI & Data:** Agent Orchestration, Generative AI, LLMs, RAG, Agentic Workflows, Vector DBs (FAISS), Embeddings, MCP, Context Engineering; Machine Learning, PyTorch, NumPy; Vercel AI SDK, LangChain/LlamaIndex, MLflow, Pydantic v2, AssemblyAI.

Education

- University of Pennsylvania**, Master of Applied Science in Computer Science Aug 2025 – May 2027
- **Relevant Coursework:** Distributed Systems, Machine Learning, DS & Algos, OS | **Cum GPA:** 3.9
- Cornell University**, BS in Food Science, minor in Business | **Major GPA:** 3.9, **minor GPA:** 4.0 Aug 2020 – May 2022
- **Relevant Coursework:** Interdisciplinary Core: full suite of Life Science & Engineering prerequisites
- China Agricultural University**, BEng in Food Science & Engineering (Dual Degree Program) Aug 2018 – Jul 2020

Experience

[Portfolio](#)

- Software Engineering Intern**, MedicScribe – Philadelphia, PA – Part-time Jul 2026 - Present
- Owned front-end and secure data-pipeline development for an EMS application piloted by **Stanford’s EMT group**, translating EMT workflows and HIPAA requirements into reliable mobile product behavior for high-stress environments.
- Software Engineering Intern**, Cronwell – New York, NY – Full-time Jul 2026 - Sep 2026
- Built and shipped systems across **AI agents, distributed job processing, WebAssembly edge workloads, serverless infrastructure, and enterprise SSO** in a three-month startup internship.
 - Designed a distributed coding-agent architecture with Redis/BullMQ and ported AI workloads to **Akamai WebAssembly edge infrastructure**; benchmarked **32 concurrent agents at zero failures and P99 3ms**.
- Project Manager**, Amerigo Scientific – Hauppauge, NY – Full-time June 2023 – May 2025
- Built and deployed **agentic marketing automation workflows in JavaScript and Python**, translating repetitive business processes into software workflows and using performance data to improve outreach scalability and pipeline efficiency.
 - Led cross-functional initiatives across Engineering and Sales; negotiated **100+ global agreements** across North America, Europe, and Asia, contributing to **85% of annual revenue**.
- Clinical Research Coordinator**, Cornell University – Ithaca, NY – Part-time Dec 2022 – Jun 2023
- Directed end-to-end study pipelines for 200 participants/quarter at 100% IRB/GLP compliance; leveraged **R** for statistical modeling and cross-functional coordination for peer-reviewed publications.
- Associate Scientist**, Eurofins Lancaster Laboratories – Lancaster, PA – Full-time Aug 2022 – Dec 2022
- Performed client sample testing in BSL-2 per cGMP/GLP/USP; collaborated with PM to ensure on-time precise delivery.
- Research & Data Analysis** | *Insect Science* co-author - Leveraged **R** & **BLAST** for meta-analysis 2019 – 2022
- Conducted large-scale database analysis on USDA nutritional records via spreadsheets to anchor core user tracking features.

Projects

[GitHub : dingonewen](#)

- WebAssembly Edge Agent Fleet** | [GitHub](#) | WebAssembly (Spin) · Akamai · StarlingMonkey Jul 2026 – Aug 2026
- Ran **32 concurrent AI agents** as self-contained Wasm (a full JS engine — StarlingMonkey — compiled to .wasm) on a serverless edge platform (Spin/Akamai) — **zero-failure concurrency at P99 3ms**, no Node/containers.
 - Each agent runs a **sequential multi-turn tool-calling loop** (LLM → tool-calls → execute, up to 20 turns) with isolated per-agent state; benchmarked **9ms** Wasm cold-start vs **<1ms** warm.
- Terminal Coding Agent** | [GitHub](#) | Mastra · Fastify · Redis/BullMQ · LibSQL · OTLP/Logfire Jul 2026 – Aug 2026
- Built an **open-source terminal coding agent** on **Mastra** — a **TypeScript monorepo** (pnpm, 5 packages) with supervisor + subagent architecture, tool-level access control, and a lazy-loaded **skills system** that cut the supervisor prompt by 60%.
 - Designed a **distributed HTTP API** on **Fastify** + **Redis/BullMQ** with **Docker Compose** orchestration — **per-conversation distributed locks** ensure same-session jobs stay sequential across workers while different sessions run in parallel.
 - Implemented **LibSQL**-backed memory (working, observational, semantic recall) with auto-compaction, **OTLP** trace export (Logfire/LangSmith) tied to session thread IDs, and runtime **multi-provider switching** (Anthropic, OpenAI, Google, DeepSeek).
- MedicScribe** | Lead Engineer | Expo/React Native · Next.js · Supabase · Claude API Jul 2026 – Present
- Designed a **HIPAA-compliant ambient scribe app** for EMTs — **Expo/React Native** + **Next.js** on **Postgres (Supabase)** with Row-Level Security, AES-256-GCM PHI encryption, and JWT auth; built an async **AI pipeline** (AssemblyAI diarization → NEMSIS reports) with degradation handling and a data-governance suite enforcing per-group retention via cron jobs.
- Hybrid Search Engine** | [GitHub](#) | FastAPI · FAISS · Sentence-Transformers · RRF Apr 2026
- Engineered a **hybrid RAG search engine**: multi-threaded **C++ BM25 server** + Python FastAPI over **FAISS** embeddings; **Reciprocal Rank Fusion** merges text + semantic signals at sub-millisecond latency.