

# Edison Xu

US Citizen | Edisonapply@gmail.com | Open to Relocation (Nationwide) | 626-552-1696  
[www.edisonwhale.com](http://www.edisonwhale.com) | [linkedin.com/in/Edisonwhale](https://linkedin.com/in/Edisonwhale) | [github.com/edisonwhale](https://github.com/edisonwhale)

## EDUCATION

|                                                          |                     |
|----------------------------------------------------------|---------------------|
| Georgia Institute of Technology                          | Aug 2022 - Dec 2024 |
| Master of Science in Computational Science & Engineering |                     |
| University of California - Santa Barbara                 | Oct 2019 - Mar 2021 |
| Bachelor of Arts, Applied Economics                      |                     |

## EXPERIENCE

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                     |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| Highmark Inc.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Aug 2024 - Present  |
| Software Engineer (AI/ML team)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Remote, USA         |
| <ul style="list-style-type: none"><li>Shipped a real-time AI call-center training platform (<b>React/FastAPI/WebSockets</b>) that composes configurable personas and scenarios into dynamic system instructions for low-latency voice roleplays, reducing new-hire ramp-up time by 28%.</li><li>Designed an <b>LLM-as-judge</b> evaluation system calibrated on human-scored transcripts for reliable coaching feedback.</li><li>Built an enterprise <b>AI-agent lifecycle control plane</b> integrating versioned registry metadata, policy gates, durable evaluation runs, GitLab promotion and observability across staging and production, standardizing agent delivery.</li><li>Architected a <b>multi-agent evaluation harness</b> using planner-executor-judge loops and structured handoffs to run owner-defined and holdout suites, catching quality, safety, latency, and cost regressions before they impacted users.</li><li>Engineered an enterprise knowledge workspace on <b>Vertex AI Search</b>, implementing a <b>RAG</b> pipeline with metadata-filtered retrieval, configurable chunking, and citation-linked answers grounded in private, access-controlled datastores.</li></ul> |                     |
| Highmark Inc.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | May 2024 – Aug 2024 |
| Software Engineer Intern (AI/ML team)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Remote, USA         |
| <ul style="list-style-type: none"><li>Developed a <b>Retrieval-Augmented Generation</b> (RAG) system using Flask, integrating embedding models and Gemini LLM to answer employee queries on internal policies, summarize documents with sources, and cut search time 70%.</li><li>Engineered an internal <b>LLM API gateway</b> exposing a single fault tolerant Gemini endpoint with automatic retries and structured request logging, standardizing GenAI access through one shared platform layer.</li><li>Built a <b>multimodal</b> AI assistant on Google Vertex AI (Gemini), enabling employees to process text, PDFs, images, and voice through a <b>React</b> and <b>FastAPI</b> stack, supporting an enterprise AI platform handling 6M+ annual prompts.</li></ul>                                                                                                                                                                                                                                                                                                                                                                                                                            |                     |

## PROJECTS

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| Conductor: Distributed Multi-Agent Workflow Runtime (Go)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Apr 2026 - June 2026 |
| <ul style="list-style-type: none"><li>Engineered a <b>Kubernetes</b>-native Go agent runtime with event-sourced durable execution and per-step checkpoints, making long-running multi-step LLM workflows crash-resumable without duplicated tool calls or wasted model spend.</li><li>Designed a distributed <b>Kafka</b>-backed multi-agent orchestration engine that schedules DAG-based LLM workflows across a scalable worker pool, using per-task retries to replay failed tasks instead of restarting the full run.</li><li>Built observability for every model and tool call with event-sourced tracing, audit logs, and real-time workflow monitoring.</li></ul> |                      |
| Engram: Cross-Session Memory MCP for AI Coding Agents (Python)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Jan 2026 - Mar 2026  |
| <ul style="list-style-type: none"><li>Built an <b>MCP</b> memory harness that hands off session state between AI agents, cutting repeated context tokens ~40%.</li><li>Engineered short- and long-term memory with <b>event-driven</b> consolidation that distills raw sessions into durable decisions, project facts, and inferred user preferences, with stale memory cleanup and conflict detection.</li><li>Validated memory reliability via <b>Recall@5/MRR</b> evals and traces to diagnose rank regressions and stale/injected context.</li></ul>                                                                                                                 |                      |
| FigBrain: Figma Multi-Agent Plugin (Python/TypeScript)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Apr 2025 - May 2025  |
| <ul style="list-style-type: none"><li>Built an MCP-powered Figma multi-agent for the Microsoft AI Hackathon using a <b>LangGraph supervisor agent</b> and ReAct tool-use loops to convert natural-language prompts into FigJam components at ~85% first-try accuracy.</li><li>Cut API costs ~28% and latency via <b>PostgreSQL</b> backed short/long-term memory, plan/result caching, and model routing.</li></ul>                                                                                                                                                                                                                                                      |                      |

## SKILLS

|              |                                                                                                 |
|--------------|-------------------------------------------------------------------------------------------------|
| Languages    | Python, TypeScript, JavaScript, Java, Go, SQL, Agentic AI Coding (Claude Code, Codex, Cursor)   |
| AI / LLM     | RAG, Multi Agent Architecture, LLM-as-Judge, Embedding Models, Fine-tuning, Harness engineering |
| Frameworks   | React, Vue, Next.js, FastAPI, Flask, Node.js, Express.js, LangChain, MCP, PyTorch, Hugging Face |
| Data / Infra | PostgreSQL/pgvector, Redis, Kafka, Docker, Kubernetes, CI/CD, AWS, GCP, OpenTelemetry           |