

Trevor Grafton

<https://trevor-grafton.vercel.app/> (portfolio: projects, demos and slides)

Greater Boston | trevorjgrafton@gmail.com

GitHub: graftonatwit | LinkedIn: Trevor Grafton

Summary

Computer Science student and Service Desk Analyst Co-op building LLM systems on Azure and local models, including a working AI phone agent for an enterprise service desk. Seeking internships for spring 2027 (January to April, or May start), part-time work for summer 2027 (May to August) and full-time roles from fall 2027 (September start).

Relevant Course Work

Systems Administration, Software Engineering, Network Programming, Wireless Networks, Algorithms, Data Structures, Network Administration, Operating Systems

Education

Wentworth Institute of Technology — Boston, MA

Bachelor of Science in Computer Science | Expected May 2027 | Dean's List (Spring 2025, Spring 2026)

Claude Code 101, Claude Academy (Anthropic) | 2026

Experience

Service Desk Analyst Co-op, BXP | May 2026 – Present

- Support end users and devices in a hybrid environment (Intune, Autopilot, Entra ID, Active Directory, Microsoft 365) through ServiceNow incidents and Rapid7.
- Built the service desk's AI voice agent (see Projects), projected to save ~\$20K/year at under \$1K/year to run.
- Deployed the voice agent on Azure: Azure OpenAI, Communication Services, App Service, Key Vault, Entra ID.
- Built onboarding and offboarding automation scripts (PowerShell) across a hybrid AD and cloud environment.
- Designed and tested Copilot Studio agents for IT knowledge management and self-service support; wrote SOPs and knowledge base articles.

IT Specialist (Part-time), BXP | Jul 2026 – Present

- Deploy, reimagine and provision Windows devices with Intune and Autopilot; manage accounts, mailbox permissions and groups across AD, Entra ID and Exchange Online.
- Evaluated AI agents for ticket intake, troubleshooting guidance and documentation search, and built knowledge-based AI solutions from internal IT documentation.

Technical Skills

AI & LLM Engineering: RAG, prompt engineering, LLM agents, tool calling, speech-to-speech voice agents, Azure OpenAI Realtime, local LLMs and embeddings, guardrails

AI Development Tools: Claude Code, GitHub Copilot, agentic coding workflows, automated code review

Languages: Python, Java, C, C++, C#, JavaScript, SQL, HTML, CSS, Bash

Cloud & DevOps: Azure (Communication Services, App Service, Key Vault, Event Grid, Entra ID), GitHub Actions, Docker

Web & Data: FastAPI, Flask, PostgreSQL (pgvector), MySQL, SQLite, REST APIs, ServiceNow Table API, Pinecone

Security: OAuth 2.0 device code flow, MFA (TOTP, Microsoft Authenticator), secrets management, RBAC

IT & Networking Tools: Active Directory, Intune, ServiceNow, Rapid7, Wireshark, iPerf3, Cisco Packet Tracer

Projects

AI Voice Agent for IT Service Desk | Jun 2026 – Present | Python, FastAPI, Azure OpenAI Realtime, ServiceNow API

- Built an AI service desk analyst that answers a real phone line and files a categorized ServiceNow incident mid-call.

- Streamed call audio over WebSockets into Azure OpenAI's Realtime API at 390 ms median turn latency (141 turns).
- The live model has only talk, transfer and hang-up tools and a tool-less model writes the ticket, so a prompt injection has nothing to reach.
- Added Entra device code caller verification and an operator console with live monitoring, TOTP MFA and RBAC.
- Measured about \$0.05 per call-minute from Azure billing; 906 automated tests.

Local AI Agent Team (JARVIS) | Oct 2026 – Present | OpenClaw, Ollama (qwen2.5:7b), Discord

- Built a self-hosted multi-agent team: a JARVIS boss agent assigns work to Research, Coder and QA agents.
- Runs on a local qwen2.5:7b model through Ollama and OpenClaw on a homelab, controlled from Discord.
- Hardened the setup: loopback-only gateway, bot token in an environment variable, Discord plugin explicitly trusted.
- Built an office GUI with a live event log (task_assigned, handoff, file_saved) to watch each task move through review.

Networked Machine Learning Algorithm | Mar 2026 | Python, Q-Learning, UCB bandit, NumPy, Matplotlib

- Built a Python network simulation where Q-learning agents tune transmission probabilities to boost throughput.
- Compared five transmission strategies (random baseline, Q-learning, adaptive step size, hybrid, UCB bandit) with a fairness term, and saved learned state across runs.
- Visualized cumulative success rates, collisions and probability evolution to evaluate learning performance.

Roblox Games: Steal a Dinosaur, Pet Cafe Tycoon, Riposte | Sep 2026 – Present | Luau, Rojo | github.com/grafonawit/pet-cafe-tycoon

- Building Roblox games in Luau with Rojo, Wally, StyLua and Selene: Steal a Dinosaur (in development) and Pet Cafe Tycoon (public repo).
- Steal a Dinosaur has 20 server services and 43 Lune spec suites.
- Riposte, a 1v1 duel game, enforces its backstep range rule with a test.

Healthcare Referral Management System | Sep 2026 – Present | Personal design project | Python, FastAPI, PostgreSQL/pgvector, local LLMs

- Designing a full-stack FastAPI system that centralizes fax, phone and email referrals in one PostgreSQL database.
- Designing a RAG pipeline with hybrid search and a reranker to automate multi-day monthly reporting.
- Planning to run the embedding model and LLM locally so patient data never leaves internal servers.

Full-Stack Restaurant Management App | Feb 2026 | Flask, MySQL, OpenCV | <https://youtu.be/mkLY7o9TAsI>

- Developed a full-stack Flask web app with complete CRUD functionality for dynamic data management.
- Integrated MySQL for persistent storage and retrieval of restaurant data.
- Designed a face-recognition login flow using Python and OpenCV.