

ANTONIO YUN

+1 (425) 412-0532 | antoniyun2000@gmail.com | Lynnwood, WA | [linkedin.com/in/antonio-yun/](https://www.linkedin.com/in/antonio-yun/)

EDUCATION

Georgia Institute of Technology

Aug 2023 - Aug 2026

Master's, Computer Science

- Specialization: Artificial Intelligence

University of Washington

Sep 2018 - Dec 2021

Bachelor's, Computer Science

EXPERIENCE

Visang Care

Jan 2026 - Present

Software Engineer Contractor

Seattle, WA

- Architected and built an internal operations application end-to-end using **Claude Code**, cutting manual operational overhead.
- Enabled scaling from 2 pilot hotels to 100+ deployments by designing reliable backend ingestion and location-aware alerting components.
- Powered real-time staff emergency alerts by developing indoor positioning algorithms using Bluetooth (BLE) device signals.
- Improved on-call triage speed by instrumenting logs and metrics and authoring runbooks covering 138 rooms across pilot sites.
- Cut issue resolution time by writing test cases and reproducing/documenting bugs for distributed real-time storage services.
- Streamlined hotel operations by building web platforms for product management, customer interaction, and admin dashboards.

Georgia Institute of Technology

Jan 2025 - Dec 2025

Research Member for Smart Stadium

Remote

- Extracted real-time player performance data by deploying **PyTorch** and **TorchVision** pipelines on CUDA GPUs across sports simulations.
- Enabled Markov chain modeling of football plays by managing cross-functional dataset development using **SAM2** and game simulations.
- Improved play-modeling accuracy by researching Madden simulation methods to analyze gameplay mechanics for strategic insights.
- Achieved precise spatial mapping of game footage by leading a team training Roboflow models for player and keypoint detection.
- Led a sub-team of 2 to 3 engineers within a 15-member ML research group, presenting findings in regular team meetings.

Basil Tech

Oct 2022 - Jun 2023

Software engineer

Seattle, WA

- Enhanced map-based application functionality by engineering React interfaces and integrating complex API calls for end users.
- Drove on-time delivery of project milestones by participating in weekly Agile planning across a 15-engineer team.
- Delivered nonprofit community applications by collaborating end-to-end on project management with engineers and stakeholders.

PROJECTS

RAG Project - [Link to project](#)

- Engineered an end-to-end Retrieval-Augmented Generation (**RAG**) backend using Python and **FastAPI**, processing PDF documents through text extraction, recursive chunking, local embeddings, semantic retrieval, and LLM-generated responses with source citations.
- Implemented PostgreSQL + pgvector vector storage and cosine-similarity search with 384-dimensional Sentence Transformer embeddings, enabling context-aware retrieval across uploaded documents without relying on hosted embedding APIs.
- Integrated a locally hosted Ollama LLM through **REST APIs** and built a modular, production-oriented architecture with **SQLAlchemy**, **Alembic** migrations, **Pydantic** validation, and a 108-test pytest suite covering ingestion, retrieval, search, and generation workflows.

SKILLS

Languages: Python, Java, C/C++, SQL, JavaScript

Backend: FastAPI, SQLAlchemy, Alembic, Pydantic, REST APIs, MySQL

AI/ML: PyTorch, TorchVision, CUDA, vector search (pgvector), SAM2, Roboflow

Tools: Git/GitHub, Linux/Unix, pytest, Docker

LLMs: Claude, Codex, Cursor, Gemma, Gemini, RAG