# Caleb Torres

Startup rockstar | Problem solver | AI agent builder

calebtorres9@protonmail.com 253-275-5757 linkedin.com/in/calebtorres1826 US citizen Greater Phoenix Area

## SUMMARY

Startup-minded Staff Software Engineer with 9+ years of experience building high-throughput infrastructure, low-latency microservices, and LLM-powered agent pipelines across startups and Microsoft. Recognized for driving key business impact, scaling 0-to-1 products, leading technical direction, and prioritizing fast, metric-driven iteration to solve core user problems.

## EXPERIENCE

**Staff Software Engineer** May 2024 – Sep 2026

Microsoft

- Led a team of 3 full-stack engineers to architect a healthcare AI agent pipeline, automating unstructured document processing and reducing time from 40 minutes to 3 minutes per document.
- Designed and built an LLM-powered semantic search pipeline featuring intent mapping and tool calling across 17k+ documents, reducing search and accurate retrieval times from 50 minutes to 2 minutes.
- Architected an Azure-based infrastructure for a healthcare Saas system.

**Senior Software Engineer** Jan 2022 – May 2024

Microsoft

- Architected end-to-end DPU infrastructure, leveraging Terraform to automate provisioning and accelerate deployment cycles.
- Designed and optimized a high-performance UI for real-time DPU management and telemetry.
- Engineered high-throughput backend microservices to deliver seamless, low-latency communication between DPUs and frontend applications.

**Full-Stack Engineer** Feb 2019 – Jan 2022

Fungible

- Helped shift the engineering team culture to remove layers between ICs and decisions.
- Built and maintained interfaces for viewing infrastructure, inspecting attached hosts, configuring resources, and tracking performance.
- Implemented the services behind those interfaces (storage, network management, telemetry, node management/log collection, and API gateways).
- Turned complex infrastructure operations into usable interfaces and reusable UI components, collaborating with designers on Data Center Composer and GPU Connect.

**Founding Full-Stack Engineer** Jan 2017 – Feb 2019

LeapAI

- Owned product development end-to-end, from brainstorming ideas, database schema design, API development, UI building, and deployment to talking to users.
- Eliminated useful but not necessary features based on metrics, reducing main feature shipment from 1 week to 3 days and increasing daily active users from 12 to 70.
- Shifted team culture to prioritize product output over hours worked to meet deadlines, improving IC ownership and helping secure VC funding.

**Software Engineer Intern** Sep 2016 – Dec 2016

Facebook

- Learned how team collaboration works in a big company.

**Software Engineer Intern** Nov 2013 – Nov 2014

Servise

- Learned that speed and user feedback matter when working at a startup.

## EDUCATION

**Master's Degree, Computer Science** 2015 – 2016

Carnegie Mellon University

**Bachelor's Degree, Computer Science** 2011 – 2013

Manhattan College

## SKILLS

| | |
|---|---|
| **Languages** | Python (Mypy, Pytest, Ruff, uv), JS/TS, Go, Rust, SQL |
| **Frameworks** | Django, SQLAlchemy, FastAPI, Flask, Node.js, React, Next.js |
| **AI** | Langchain, OpenAI, Anthropic, Gemini, etc. |
| **Databases** | Postgres, MySQL, MongoDB, SQLite, Redis, Cassandra, Spark |
| **Cloud** | AWS (EC2, ECS, EKS, Lambda, S3, RDS, DynamoDB, etc.), Azure, GCP, Docker, Kubernetes, Terraform, OpenTofu, Linux |
| **Architecture** | Monolithic, Microservices, Distributed, Event-driven |
| **API design** | RESTful, GraphQL, gRPC, OpenAPI |