

# Gowtham Arulmozhi

541-286-2416 • gowtham.arulmozhi@gmail.com • linkedin.com/in/gowtham-arulmozhi • github.com/wothmag07 • gowthamarul.com

## EDUCATION

### Oregon State University

Sep 2023 - Jun 2025

Master of Science in Artificial Intelligence (GPA: 3.76/4.0)

Corvallis, OR

- **GTA, CS 479 (Open Source Projects)** — Mentored 200 students, streamlined course structure, and resolved team blockers.
- **Certifications:** AWS Solutions Architect Associate, Certified Kubernetes Administrator, AWS Cloud Practitioner, HashiCorp Terraform Associate

### Anna University

Jul 2017 - May 2021

Bachelor of Engineering in Electronics and Communications Engineering (GPA: 8.3/10)

India

## WORK EXPERIENCE

### Beaverhand Inc.

Sep 2025 - Present

AI Engineer

Remote, US

- Led 0-to-1 architecture and development of **AI-native recruitment platform** (React/TypeScript, Supabase, FastAPI) automating candidate screening, live AI interviews, and structured evaluation workflows, replacing manual first-round interviews.
- Architected **real-time AI interview engine** (phone + technical) using Gemini Live native audio and ElevenLabs ConvAI, with Mem0-backed session memory and dynamic grounding from resumes and job descriptions for rubric-driven technical and behavioral evaluation.
- Designed scalable **multi-modal recording and evaluation pipeline** (screen, webcam, audio) with GCS resumable uploads and VLM-based screen analysis, powered by an async multi-agent backend (LlamaIndex, Gemini 2.5 Flash) orchestrated via Celery/Redis.
- Deployed production-ready infrastructure on **GCP** using Cloud Run containers, Docker multi-stage builds, nginx reverse proxy to ML inference services, and Supabase Edge Functions, enabling low-latency streaming and high-volume media ingestion.

### Oregon State University

Mar 2024 - Dec 2024

Data Scientist

Corvallis, OR

- Built **schema mapping system** for NIH Translator project using Word2Vec (Gensim) for semantic column naming and LLMs to generate SQL views with joins, type casting, and aggregations—enabling natural language-based, schema-agnostic access to biomedical databases.
- Designed benchmarking framework to evaluate SQL outputs from LLMs, achieving **82% precision** and **73% recall** in generating valid and semantically correct SQL queries compared to expert-curated ground-truth queries.

### National Payments Corporation of India

Aug 2021 - Sep 2023

Software Engineer

India

- Built XGBoost-based **fraud detection pipeline** on KubeFlow to identify Aadhaar transaction anomalies (e.g., geographic, velocity), achieving real-time detection refined for **high recall (0.74)** and **strong precision (0.81)**.
- Prototyped YOLOv5-based **license plate recognition system** to replace RFID for NETC FASTag, achieving **0.87 mAP@0.5**, **92% OCR accuracy**, and **25 FPS** real-time throughput.
- Architected **fault-tolerant, cross-site high-availability platform** for NETC FASTag processing **10M+ daily transactions** using HAProxy, Kafka, Cassandra, KeyDB, and the ELK stack, ensuring **zero transaction loss and seamless failover**.
- Developed **observability platform** for the Central Bank Digital Currency (CBDC) application, leveraging Prometheus, Grafana, and EFK, which slashed incident response times **from 3 hours to under 5 minutes**.

## PROJECTS

### Doc-Vault MCP (MCP HF Hackathon) [GitHub] [Live Demo]

Nov 2025 - Dec 2025

- Built **Doc-Vault MCP** from 0→1, an AI-native documentation platform with markdown editing, wikilinks, full-text search (SQLite FTS5 + BM25), and vector RAG retrieval (LlamaIndex + Gemini), including dual LLM integration (MCP tools + Gemini chat), deployed on HuggingFace Spaces.
- Architected **secure multi-tenant system** with HuggingFace OAuth, JWT authorization, and isolated per-user vaults, enabling MCP agents to read/write/search notes with automatic indexing, real-time sync, and knowledge-graph mapping.

### Amazon Electronics Product Assistant [GitHub] [Live]

Jul 2025 - Nov 2025

- Built a **self-correcting RAG agent** using LangGraph with document grading and automatic query rewriting, indexing **75K+ review chunks** from the Amazon Reviews 2023 dataset to generate grounded, citation-backed product recommendations.
- Implemented **hybrid retrieval** (vector + BM25) with a **6-metric evaluation suite**, achieving 0.80 citation grounding accuracy and 0.75 retrieval relevance. Deployed on **GCP Cloud Run** with SSE-based token streaming. (*FastAPI, Supabase, Next.js, Docker, GCP*)

### Real-Time AI Video Avatar System (in collaboration with Beaverhand) [GitHub] [Demo]

Dec 2024 - Jun 2025

- Engineered **multimodal AI avatar system** that converts text and speech to synchronized video using GeneFace++ and Google STT & TTS, achieving **sub-2-minute processing time** validated across **25 test users**.
- Deployed scalable serving pipeline with **FastAPI** and **Docker**, exposing a REST API for asynchronous avatar generation with consistent real-time performance. (*Tools: Python, PyTorch, OpenCV, FastAPI, JavaScript, LangChain, Docker*)

## TECHNICAL SKILLS

**Languages:** Python, SQL, C/C++, JavaScript (Reactjs, Nodejs), Typescript, Bash

**AI Tools & Frameworks:** PyTorch, TensorFlow, scikit-learn, HuggingFace, JAX, FastAPI, Pydantic, Llama-index, LangChain, LangGraph, LangSmith, Langfuse, MCP, Google ADK, AsyncIO, Qdrant, vLLM, ONNX Runtime, CrewAI, Gradio, A2A, Tableau, PowerBI, pytest

**Data & ML Ops:** Airflow, PySpark, Kafka, dbt, MLflow, Databricks, Docker, Kubernetes, Prometheus, Grafana, Jenkins, Github Actions

**Databases:** PostgreSQL, MySQL, Snowflake, MongoDB, Redis, Elasticsearch, Cassandra

**Cloud:** GCP (Vertex AI, Cloud Run, Dataflow, Pub/Sub), AWS (S3, Lambda, Bedrock, SageMaker), Azure (OpenAI, Cosmos DB, AKS)