Yashwant Gandham

Boulder, CO | gyashwant2002@gmail.com | +1 303-414-8260 | LinkedIn | Github

Technologies

Software & Backend: Python, FastAPI, PostgreSQL, Redis, REST API

AI Systems: PyTorch, Pinecone, LangGraph, Neo4j

Infrastructure & Performance: Git, GCP, Docker, Railway, Vercel, Linux, Perfetto, Nsight

Experience

Backend Infrastructure Engineer NovaChat AI – Boulder, CO Aug 2025 – Present

- Built conversational AI marketing agent with LangGraph and FastAPI achieving **18%** response rate across **800+** messages, generating **400+** platform signups, **$50K**/month revenue growth, and **2** major client onboardings.
- Collaborated with customers to scope and deliver **5+** high-impact features including onboarding optimization (**70%** time reduction), native media sync (**90%** faster), and **24/7** auto-response system (**58%** latency improvement), directly supporting **$50K**/month revenue growth.
- Developed and maintained backend APIs for response management and vector search operations, serving as core infrastructure for multi-tenant conversation pipelines.
- Coordinated multi-platform deployment across Railway (backend) and Vercel (frontend) using **Git** release workflows for production feature updates.

Machine Learning Systems Researcher Boulder AI & Infra Lab – Boulder, CO Sept 2025 – Present

- **Profiled 3.3B** parameter Pi0 VLA model using **Perfetto**, identified vision encoder bottleneck (sequential processing of cameras) **(50ms per camera)**.
- Designing **parallelization strategy** to optimize multi-camera inference pipeline **toward 73ms** latency target.

Research Assistant (AI & Robotics), HIRO Lab – Boulder, CO Oct 2024 – Aug 2025

- Built **computer vision** pipeline processing Tobii Pro eye-tracking data using **SAM, DINO, and CLIP** for attention-based object segmentation, generating multiple visualization outputs from **pupil density heatmaps.**
- Implemented reinforcement learning agents **(DQN, PPO, A2C)** for robotic block assembly in **OpenAI Gym** environments, achieving **>85%** task success validated across **500+ episodes.**

Projects

DocChat: GraphRAG System with Self-Correction 2026
Python|Docker|FastAPI|Pinecone|GCP|Langraph|VertexAI|Neo4j

- Engineered a Hybrid GraphRAG system combining semantic retrieval and graph-based relationship reasoning, improving multi-hop query accuracy by **3.4×** across **150+** research papers.
- Implemented an agentic self-correction loop that evaluates responses and retriggers retrieval when confidence is low, reducing hallucinations by **25%**.
- Built a distributed document extraction and indexing pipeline using GROBID and custom parsers, achieving **99%** uptime and reducing LLM inference costs by **40%** through topology-aware retrieval.
- Built an interactive React Mind Map to visualize citation networks across **10K+** nodes.

ML Observability & Drift Monitoring Platform 2025
Python|FastAPI|Docker|PostgresSQL|GCP|SQLAlchemy

- Designed and implemented an ML observability platform ingesting **3000+** predictions/day across multiple binary classifiers, logging predictions, ground-truth labels, and **15+** input features into PostgreSQL; deployed containerized FastAPI on GCP Compute Engine with Docker/systemd achieving **<100** ms latency for prediction and ingestion.
- Built a batch analytics pipeline processing **3000+** prediction logs/day to compute rolling accuracy, precision, recall, F1-score, and latency percentiles, aggregating **7** key metrics per model into optimized summary tables.

Education

Master of Science in Data Science | University of Colorado Boulder Aug 2024 – May 2026
Data Center Scaling|Modern LLM|Deep Learning|Distributed Systems|Statistical methods Boulder, CO

Bachelor of Technology in Electrical Engineering Mar 2019 – Jun 2023
Jawaharlal Nehru Technological University Hyderabad, India