OpenAI
Tier 3
Builder

RAHUL GUPTA

Machine Learning Engineer

+1 (619) 764 8386 • rahul.gupta2608@yahoo.com • https://rahulg.info • San Diego, CA

ANTHROP\C
TIER 2 BUILDER

Experience

Fleuxlabs.ai — San Diego, CA
Founder — 08/2022 - Present

LLM Platform

- Live **NextJS** LLM platform averaging over **900,000 tokens** a day, featuring multi-LLM chat integration with **GPT-4/VO**, **Anthropic**, and Perplexity.
- Platform uses Pinecone, a **vectorized** database and **Cohere** embeddings to ensure seamless **context sharing within the same chat.**
- Includes **5 attachments** per message and TTS backed by Eleven labs.
- **Actively working** on Puppeteer based Dockerized **agentic control systems** with multi-llm access running on proxies in a controlled access space.

Walmart Global Tech — Vancouver, BC
Software Development Engineer II — 06/2021 - 08/2022

Release Engineering

- Dockerized WalmartUS iOS and Android mobile **release pipeline**, collaborating with NX founder, Victor Savkin to create a custom monorepo.
- Implemented dependency caching and dependency graphs, reducing average **build times by 47%.**
- Subsequent builds after initial failure were **78% faster**, down to **12 minutes from 54 minutes**, enhancing developer commit velocity.

@WalmartLabs — Sunnyvale, CA
Software Development Engineer In Test — 06/2019 - 05/2021

Automation Frameworks

- Developed a ReactJS tool that **automated the entire order lifecycle** using mock APIs, enabling same-day testing and deployment across US, UK, and India.
- Eliminated production hot-fixes, significantly reducing risk and improving software quality.
- **Saved 200+ hours** of US resources and facilitated the US testing department's transition to India due to the tool's reliability and minimal maintenance requirements.

Projects

Speaker Recognition — San Diego State University — 08/2023

Natural Language Processing

- Identify speakers from King Corpus using **20ms frames.**
- Utilized **GRUs** to handle sequential data.
- Employed **Long Short-Term Memory** networks to better retain and utilize long-term dependencies.
- Implemented **batch normalizations**, dropout layers, and L2 regularization to mitigate overfitting.

Verte — San Diego State University — 11/2023

Mistral-7B Fine-tuning

- Enhanced Mistral-7B for spine-related queries using **Quantized Low-Rank Adaptation**.
- Leveraged **OpenAI Whisper** for transcription and GPT-4 Turbo for generating diverse Q&A pairs.
- Achieved at least **50% improvement** in accuracy and efficiency for **spine-related medical query responses.**

Stratas — San Diego, CA — 01/2023 - 01/2024

Computer Vision

- Built a **React Native app** integrating **speed sensors** & Bluetooth device connectivity.
- Developed a Raspberry Pi **prototype** with **OpenCV** libraries and a 4K camera for ride monitoring.
- Made a **data-driven decision** to halt further development due to scaling challenges and **universal chip shortages.**

Skills

PyTorch · Jupyter · Python · OpenAI-SDK · Anthropic-SDK · Puppeteer · AWS · Typescript · Vercel · NextJS · ReactJS · Java · Pinecone DB · Docker · TailwindCSS · Jenkins · HuggingFace · Tensor flow · Redis · Lang-chain.

Education

San Diego State University — San Diego, CA
MS Computer Science — 08/2022 - 05/2024

San Diego State University — San Diego, CA
BS Computer Science — 08/2014 - 08/2018