Piyush Choudhari

Pune, Maharashtra, India

+91 9168088565 | GitHub | LinkedIn | Portfolio | choudhari.piyush@gmail.com

AI & Backend Engineer

Technical Skills

- **Languages:** Python, TypeScript, SQL
- **Backend & APIs:** FastAPI, Flask, Django, Node.js
- **Databases:** PostgreSQL, MongoDB, MySQL, Neo4j
- **Cloud & DevOps:** AWS (ECS, EC2, RDS, S3, ECR, Fargate), Docker, GitHub Actions, GCP (Cloud SQL), Nginx

Professional Experience

Ronin Labs - AI Engineer Intern — **Jan 2025 - Jul 2025**

Multimodal Content Generation System (Agent-Based Architecture)

- Architected a modular media generation backend using **LangGraph and OpenAI APIs**, enabling a central orchestration agent to dynamically route tasks across image, video, and audio generation pipelines.
- Designed image outpainting and depth-effect video pipelines using **SDXL and ControlNet models**; developed music generation modules with **MusicGen**.
- Optimized the backend processing pipeline by supporting concurrent GPU task execution, integrating state-managed workflow behind a Flask API, **reducing task processing times to <20 seconds**.

Sketchbot - Link

- Engineered an image generation backend using FastAPI, SDXL Turbo, ControlNets, and OpenCV, generating line-art images and optimizing G-code for Dexarm sketching, **reducing draw times by 70% from 10 mins to <3 mins**.
- Integrated MongoDB for user management and S3 for asset storage, **delivering <200ms average API response times under load**.
- Deployed on AWS (g5.2xlarge GPU instances); optimized servers through consolidation, **reducing infra costs by 30% per region**.
- Implemented secure device authentication restricting access to authorized Raspberry Pi units. Automated AMI/EBS snapshot creation for rapid **(<15 min) disaster recovery and redeployments, cutting downtime by 95%.**

OnePlus 13s Quest Quiz *(Client Project)* - Link

- Collaborated in a 6-member team to develop scalable backend APIs using Node.js, TypeScript, and TypeORM, **supporting 150K+ active users with <120ms API response times** in a global quiz experience.
- Built a **Django-based admin panel** for non-technical campaign operators to manage leaderboards and user analytics.
- Managed **GCP Cloud SQL** infrastructure, ensuring high availability and reliability under peak traffic conditions for a production-grade consumer application.

Projects

KnowFlow - GitHub | Demo Video | HLD ***(Python, FastAPI, Docker, AWS, ECS, Fargate, S3, PostgreSQL)***

- Architected a document retrieval system combining PGVector embeddings with Neo4j nodes for multi-hop reasoning and structured querying via Gemini APIs, **delivering Recall@5 score of ~75%**.
- Engineered a query decomposition pipeline, optimizing chunk relevance via automated feedback loops, **increasing retrieval accuracy (Recall@5) by 10%**.
- Designed secure multi-tenant ingestion pipelines, with per-user isolation and **scalable indexing via AWS S3**.
- Developed a scalable backend (FastAPI, PostgreSQL, Neo4j, S3), containerized via Docker and deployed on AWS ECS Fargate, **achieving average latency <5 seconds for heavy retrieval workloads**.
- Setup CI/CD pipelines (GitHub Actions) automating daily builds and ECS deployments, **reducing release times to <5 minutes**. Integrated CloudWatch for logs and monitoring.

TrackML - GitHub | Demo Video ***(Python, Flask, AWS, EC2, PostgreSQL, Gemini, Groq)***

- Developed a full-stack ML model tracking platform using **Python, Flask, and AWS EC2**, integrating semantic search, metadata summarization (RAG), and automated ingestion from academic papers and websites, streamlining model comparison and tracking workflows.
- Engineered a real-time metadata parsing pipeline using **Gemini API and Groq-hosted LLaMA-4-17B** models to process research URLs, **reducing manual documentation effort by over 80%**.
- Implemented FAISS vector search with bge-small-en-v1.5 embeddings over Neon-hosted PostgreSQL, **achieving sub-200ms semantic query latency across 50+ tracked ML models**.
- Deployed backend behind Nginx reverse proxy with GZIP compression, maintaining **<300ms API response times under 100+ concurrent queries**, monitored via Nginx logs and custom API instrumentation.

Education

D.Y.Patil College Of Engineering, Akurdi — ***2022-2026***

Bachelor of Engineering A.I.D.S(CGPA of 8.12) — *Pune, Maharashtra, India*