Pradyumna Yadav

AI Engineer : Agentic Systems · AI Infrastructure · Voice AI
pradyumna.aky@gmail.com **9893569895** **Portfolio**

Education

IIIT Naya Raipur - B.Tech, Electronics & Communication Engineering 2023 CGPA: 8.25

Experience

- AI Engineer, ***FuturePath AI*** **Nov 2024 – Present**
  Products deployed across Fortune 500 enterprise clients
  - Led AI engineering on Voice Copilot, a real-time voice assistant for enterprise IT operations integrating Cisco and Genesys telephony; **Cisco onboarded as a technology partner**
  - Built LiveKit audio streaming layer with configurable STT provider for low-latency transcription and post-call AI summarization
  - Built a shared RAG pipeline across Voice Copilot and Knowledge Assist, ingesting from SharePoint and ServiceNow (PgVector, LlamaIndex), with async background workers for continuous knowledge base ingestion and delta sync
  - Built Knowledge Assist features: document clustering by topic, ticket cluster–KB article matching with user-triggered KB generation for clusters with no existing article, and language coverage tracking via embedding similarity
  - Hardened security across services (Snyk, Chainguard), managed Prisma migrations, and owned Docker/Kubernetes/GitHub Actions CI/CD
- AI Engineer, ***Graymatics*** **Apr 2023 – Nov 2024**
  - Architected GPU-accelerated multi-stream video inference pipelines on NVIDIA Deepstream, serving YOLOv7 via Triton Inference Server backed by TensorRT engines; achieved **117 QPS at 85% mAP**, directly contributing to a **$50K deal**
  - Applied FP16 QAT on YOLOv7 for ∼2x inference speedup over FP32 baseline; exported to ONNX and compiled with TensorRT for deployment across edge (Jetson Nano) and cloud (AWS)
  - Streamed frame-level metadata across 40+ concurrent video streams using Apache Kafka
- Data Science Intern, ***Ensuredit*** **Apr 2022 – Aug 2022**
  - Fine-tuned LayoutLM for structured extraction from insurance policy documents; built Streamlit demo for internal validation
  - Implemented rPPG (remote photoplethysmography) health monitoring using MTTS-CAN for blood volume pulse prediction from facial video; achieved **BP MAE of 10 mmHg** and **HR MAE of 3 BPM**
- Student Research Intern, ***NIT Trichy*** **Oct 2021 – Feb 2022**
  - Built GNN-based model for travel-time prediction on road network graphs
  - Designed autoencoder-based feature extraction pipeline on traffic datasets for downstream prediction tasks

Technical Projects

- **OpenClaw SaaS** *(In Progress)* — Managed SaaS platform on AWS Fargate; orchestrates AI agents across WhatsApp, Telegram, and Discord with a container-per-tenant ECS architecture
- **LLM Gateway** *(Planned)* — High-throughput LLM proxy for large-scale deployments; multi-provider routing, Redis rate limiting, and observability built on FastAPI/AsyncIO
- **Virtual Try-On** *(2024)* — Implemented Virtual Try-On using diffusion model checkpoints and CLIP prompt tuning; matched competitive baseline results

Technical Skills

Languages: Python, TypeScript, C++, CUDA

Frameworks: FastAPI, Celery, Prisma

AI Frameworks: PyTorch, LlamaIndex, LangChain, LangGraph, DSPy, LiteLLM | *Voice:* LiveKit

LLM Providers: OpenAI, Azure OpenAI, AWS Bedrock, Vertex AI

Computer Vision: NVIDIA Deepstream

AI Inference: Triton Inference Server, TensorRT, ONNX

Databases: PostgreSQL, pgvector, Qdrant, Redis

DevOps: Docker, Kubernetes (EKS / GKE / AKS), Nginx, AWS, GCP, Azure, GitHub Actions