Ayush Kumar

Phagwara, Punjab
+91-9955025515 ayushk1503@gmail.com linkedin.com/in/ayushkumar15/ github.com/ayushk1233

Work Experience

Applied AI Engineer Intern **Oct 2025 - Present**
ATG/BANAO Technologies

- **InterviewGod.ai Architecture & Proctoring:** Built a multilingual AI interview platform utilizing **Sarvam API** for automated call screening, transcription, and LLM evaluation, optimizing prompt structures to reduce token consumption by 40%. Engineered custom Computer Vision pipelines to detect on-screen phones and analyze lip-sync dynamics, neutralizing proxy/deepfake attempts with <2% false positives.
- **Vidya AI Voice Agent & RAG Implementation:** Engineered a stateful, low-latency conversational tutor integrating the **Gemini Live API** and **LangGraph**. Achieved sub-800ms voice-to-voice latency by optimizing WebSocket streams, and implemented a **Pinecone** vector database to enable cross-session memory and high-speed context retrieval.
- **Coding Assessment ML Pipeline:** Curated and validated a custom dataset of 5,000+ coding challenges, designing the backend retrieval infrastructure via **Pinecone** to ensure the automated agent delivered hallucination-free code evaluations.
- **Serverless Deployment & CI/CD:** Architected a highly available microservices backend using **FastAPI** and **Docker**. Deployed across a serverless **AWS** ecosystem (**ECS Fargate, ECR, Lambda**), enabling auto-scaling that seamlessly handled 3x peak traffic spikes. Automated pipelines via **GitHub Actions**, cutting deployment cycles by 25%.
- **Sales Agent & Compute Optimization:** Repurposed a Meet Recorder service into an autonomous Sales Agent pipeline. Strategically offloaded heavy GPU and ML inference workloads to **RunPod** instances, significantly reducing cloud compute overhead compared to native AWS deployments while maintaining high throughput.

AI & Decentralized Systems Intern **Aug 2025 - Sept 2025**
BlockseBlock

- **Local LLM Deployment & Hardware Optimization:** Engineered decentralized AI applications by running **Ollama** locally, deploying and optimizing **LLaMA** models to maximize inference speed and minimize memory usage under strict edge-hardware constraints.
- **Parameter-Efficient Fine-Tuning (PEFT):** Fine-tuned open-source LLMs utilizing **LoRA** and **QLoRA** techniques, significantly reducing training compute requirements while preserving model accuracy. Successfully hosted and maintained the custom model weights on **Hugging Face**.
- **Prompt Engineering & Task Orchestration:** Architected automated workflows using **LangChain** and applied advanced prompt engineering to force strict, structured data outputs (e.g., JSON validation) from the LLMs, improving decision-making reliability and task routing efficiency by 35%.

Projects

Competitor Intelligence Monitor |*Kubernetes, AWS, LangGraph, Celery, RAG, Prometheus, Grafana* **Feb 2026- Apr 2026**

- **Autonomous Market Intelligence & AI Reasoning:** Architected a distributed AI platform using LangGraph, Gemini 1.5, and ChromaDB (RAG) to autonomously monitor and compare tech competitors. Orchestrated multi-domain web scraping pipelines processing ~30K tokens/run to extract strategic market signals (pricing shifts, hiring trends), track long-term strategic drift, and generate automated cross-competitor sales battlecards.
- **Cloud Infrastructure & LLMOps Observability:** Deployed a highly available, event-driven backend on AWS EKS via Terraform, managing asynchronous workloads with FastAPI, Celery, and Redis. Engineered a production-grade LLMOps observability stack using Prometheus and Grafana, tracking p95 pipeline latency (~70s) and reducing system debugging time by ~60% while enforcing CI/CD prompt evaluation guardrails**.**
- Github Repository Link: **https://github.com/ayushk1233/Competitor-Intelligence-Monitor.git**

End-to-End MLOps Deployment: | *FastAPI, Terraform, Ansible ,Docker, Nagios, AWS* **Jun 2025 - Jul 2025**

- Built an end-to-end MLOps pipeline with Terraform, Ansible, Docker, and FastAPI, deploying a production-ready ML service on a multi-node AWS setup with Nagios monitoring.
- Automated model training, retraining, deployment, and CI/CD workflows using GitHub Actions for reliable and repeatable releases.
- Github Repository Link: **github.com/ayushk1233/mlops-house-price-deployment.git**

Achievements & Open Source

- **Aden Hive (Core Contributor):** Engineered MCP tool integrations for the agent runtime harness, enabling secure API interactions with built-in state persistence.
- **Bindu (Core Contributor):** Architected the DSPy adapter and expanded gRPC test coverage to facilitate secure Agent-to-Agent (A2A) communication via DIDs.

Technical Skills

Languages: Python, C++, SQL
AI & Machine Learning: PyTorch, TensorFlow, Hugging Face, LangGraph, LoRA/QLoRA, ChromaDB (RAG), MLflow, Scikit-Learn
Backend & Distributed Systems: FastAPI, Celery, Redis, PostgreSQL, WebSockets, REST APIs
Cloud, DevOps & Observability: AWS(EKS, ECS, Lambda), Kubernetes, Docker, Terraform, Ansible, Prometheus, Grafana, Github Actions

Education

Lovely Professional University **2022 – 2026**
Computer Science and Engineering — CGPA: 7.72 *Phagwara,Punjab*
Delhi Public School **2020 – 2021**
12th with Science — Percentage: 93.00% *Bokaro, Jharkhand*
The Pentecostal Assembly School **2018 – 2019**
10th with Science — Percentage: 94.4% *Bokaro, Jharkhand*