Ansh Singhal

Greater Noida, India | **anshsinghal3107@gmail.com** ↗ | +91-8929554991 | **anshsinghal.dev** ↗ |
linkedin.com/in/anshhh-singhal ↗ | **github.com/AnshSinghal** ↗

Summary

AI/ML Engineer specializing in end-to-end AI deployment, MLOps, and cutting-edge research implementation. Experienced in deep learning, NLP, and computer vision, with a proven track record of building scalable AI solutions

Skills

- **Programming Languages**: Python, Java, C++, SQL
- **Machine Learning & AI**: PyTorch, TensorFlow, Hugging Face Transformers, LangChain, LlamaIndex, LLMs
- **Backend & Web Frameworks**: Spring Boot, Flask, FastAPI, Django, REST APIs, gRPC, Bootstrap
- **DevOps & MLOps Tools**: Docker, Kubernetes, Redis, Kafka, Celery, Git, Prometheus, Grafana, CI/CD
- **Cloud Platforms**: Google Cloud Platform (GCP), Microsoft Azure, Amazon Web Services (AWS)

Education

- **Bennett University, The Times Group** — Greater Noida, India
  B.Tech in Computer Science (AI Specialization); CGPA: 9.58 — *2023 – 2027*

Experience

- **IntelOwl – Threat Intelligence Platform** ↗ — Remote
  Open Source Developer — *Jan 2025 – Present*
  - **Security Analysis Enhancement**: Implemented 4 new security analyzers, expanding threat detection capabilities by 23%. Built a scalable security analysis platform using React, Django, Celery, Docker, and Redis, handling 500+ concurrent analysis requests. Implemented unit tests for analyzers using pytest to validate analyzer outputs, error handling,.
  - **Infrastructure Optimization**: Fixed Django API migrations, achieving 99.9% migration success rate and reducing deployment downtime
  - **Technical Architecture**: Developed and maintained a **Django** backend exposing RESTful endpoints for OSINT enrichment, enabling easy integration.

Projects

- **DHARA: Agentic RAG — Legal Research Engine** ↗:
  - Built an agentic Retrieval-Augmented Generation system using **FastAPI, LangGraph, and Pinecone**, enabling domain-specific legal search across **10,000+ structured and unstructured documents** with **97% retrieval accuracy**. Built with prompt engineering, multi-agentic orchestration, and AI observability to improve relevance.
  - Designed a **hybrid dense + sparse retrieval pipeline** with **two-stage re-ranking (BGE + MiniLM)**, reducing irrelevant document retrieval and achieving **P95 query latency under 500 ms**.
  - Deployed containerized microservices on **AWS ECS**, supporting concurrent queries with deterministic responses and production-grade request isolation.
- **Realtime VoiceOps AI — Streaming Incident Intelligence Platform** ↗:
  - Built a **realtime voice streaming backend** using **WebSockets and Kafka**, processing continuous audio streams with **sub-200 ms per-chunk latency** and supporting **100+ concurrent live sessions** without blocking.
  - Designed an **event-driven data layer** combining **PostgreSQL (transactional incident timelines), MongoDB (raw transcripts), and Redis (session state, rate limiting)**, ensuring ordered event processing and low-latency reads under sustained load.
  - Deployed containerized ingestion and processing services on **Kubernetes**, scaling consumer workers via **Kafka consumer groups and HPA**, and implemented observability to track **end-to-end latency, consumer lag, and failure rates** in production-like environments.
- **Flood-GAN: Physics-Aware Flood Mapping** ↗:
  - Developed a **physics-informed conditional GAN in PyTorch Lightning** to translate **Sentinel-1 SAR imagery into Sentinel-2 optical outputs**, achieving **PSNR 31.25 and SSIM 0.94** on held-out flood scenes.
  - Implemented a **dual-head U-Net architecture** producing RGB imagery and water segmentation simultaneously, enforcing **NDWI-consistent supervision** and improving flood boundary accuracy under cloud-covered conditions.
  - Integrated **cloud-masked perceptual loss (VGG19)**, **speckle-preservation loss**, and **EMA-stabilized training**, ensuring consistent visual fidelity across diverse geographies.
- **Talent Insider: Scalable Recruitment Platform** ↗:
  - Built a full-stack recruitment platform using **Java Spring Boot and MySQL**, supporting **1,000+ registered users** with role-based access control and real-time job and application workflows.
  - Designed REST APIs with **bcrypt-based authentication and transactional SQL operations**, ensuring data consistency across concurrent job postings and candidate actions.
  - Optimized database queries and API execution paths to achieve **sub-100 ms response times** for core read and write endpoints under simulated load.

Leadership & Campus Roles

- **Board Member Rotary International – Youth Advisory Council** — USA
  ∘ *Sole Youth Advisory Council representative from Asia among 10+ countries.* — *Jul 2023 – Jun 2024*
- **Co-Organizer, Google Developer Groups (GDG)** — Bennett University
  ∘*Lead 12+ technical events engaging 4,000+ participants, driving 40% membership growth* — *2025 – 2026*
- **Joint Secretary, SPARK, Entrepreneurship Cell** — Bennett University
  ∘*Drove campus-wide entrepreneurship awareness, contributing to the launch of 15+ student startups* — *2025*