Ansh Singhal

+91-8929554991 | anshsinghal3107@gmail.com | linkedin.com/in/anshhh-singhal | github.com/AnshSinghal | anshsinghal.dev

Professional Summary

Backend and AI/ML engineer with hands-on experience in agentic AI, secure LLM infrastructure, DevSecOps, and cloud-native deployment. Built production systems for RAG, real-time streaming, and AI security enforcement with measurable impact across latency, reliability, and safety. Open-source contributor with shipped work in threat intelligence, automated testing, and platform stability.

Experience

Development Trainee Feb 2026 – Present
CyberUltron Consulting Pvt. Ltd. *Remote*

- Built a secure FastAPI-based LLM API gateway with Redis-backed authentication, SHA-256 API keys, model allowlists, and centralized pre and post LLM security enforcement.
- Implemented guardrails for PII detection and prompt-injection mitigation, reducing unsafe LLM behavior by 90%+ across protected endpoints.
- Designed observability and audit telemetry across 50+ endpoints using PostgreSQL and Redis for traceability, monitoring, and risk reporting.
- Supported high-throughput inference workloads with enforcement-ready routing for 1000+ RPS.

Open Source Developer Jan 2025 – Present
IntelOwl Threat Intelligence Platform *Remote*

- Contributed 4 production security analyzers, expanding platform threat detection coverage by 23%.
- Added pytest-based validation suites to improve analyzer reliability and regression protection.
- Fixed Django migration and data consistency issues, improving migration success to 99.9%.
- Resolved Docker networking and caching issues to improve deployment stability and analyzer execution.

Projects

DHARA – Agentic RAG Legal Research Engine | *FastAPI, LangGraph, Vector Search* Github Link

- Built an agentic RAG system over 10,000+ legal documents with 97% retrieval accuracy.
- Achieved sub-500ms P95 latency with citation-traced answers on AWS containerized microservices.
- Implemented multi-agent query decomposition, reranking, and deterministic grounding for legal search.

Realtime VoiceOps AI | *Kafka, WebSockets, Kubernetes* Github Link

- Built a real-time voice streaming backend with sub-200ms latency for 100+ concurrent sessions.
- Scaled processing with Kafka consumer groups and Kubernetes HPA backed by PostgreSQL, MongoDB, and Redis.
- Tracked consumer lag and P95 latency to monitor performance across distributed streaming pipelines.

Flood-GAN: Physics-Aware Flood Mapping | *PyTorch Lightning, Computer Vision* Github Link

- Developed a physics-informed GAN for SAR-to-optical translation, achieving PSNR 31.25 and SSIM 0.94.
- Used a dual-head U-Net with NDWI supervision to improve flood-boundary detection under cloud cover.
- Stabilized training with EMA updates and speckle-preservation loss for consistent visual fidelity.

Technical Skills

Languages: Python, Java, C++, SQL, JavaScript

AI & ML: LLMs, Agentic RAG, LangChain, LlamaIndex, Hugging Face, PyTorch, TensorFlow, spaCy, NLP Pipelines

Backend: FastAPI, Django, DRF, Flask, Spring Boot, Celery, JWT Auth, Django Channels, REST APIs

Security: Presidio, llm-guard, OPA, Envoy, OWASP LLM, Prompt Injection Mitigation, PII Detection

Infrastructure: Docker, Kubernetes, Kafka, RabbitMQ, Nginx, CI/CD, Prometheus, Grafana

Cloud: AWS, GCP, Azure

Databases: PostgreSQL, Redis, SQLite, Pinecone, Chroma, Milvus

Tools: Git, Linux, LiteLLM, Uvicorn, ASGI

Education

Bennett University, The Times Group Greater Noida, India
B.Tech in Computer Science (AI Specialization), CGPA: 9.47 *Aug. 2023 – May 2027*

Leadership & Impact

- **Youth Advisory Council, Rotary International** — Sole Asia representative, global youth initiatives
- **Co-Organizer, GDG Bennett University** — Led 12+ events, 4,000+ participants, 40% growth
- **Joint Secretary, Entrepreneurship Cell** — Supported launch of 15+ student startups