GAURAV CHAUHAN

Mumbai, Maharashtra | (+91) 9930822591 | gauravc2708@gmail.com

2796gaurav.github.io | linkedin.com/in/gauravc2708
https://medium.com/@gauravc2708 | https://github.com/2796gaurav

PROFESSIONAL SUMMARY

Senior ML Engineering Leader with 8+ years architecting production AI/ML systems in fintech and insurance. Expert in LLM deployment, MLOps infrastructure, AI governance, and regulatory compliance. Led cross-functional teams building scalable GenAI solutions serving 1M+ users, $108K annual cost savings and 40% operational efficiency gains. Proven track record implementing enterprise-grade guardrails, model monitoring systems, and establishing compliance-first ML practices for regulated industries on AWS/GCP platforms.

CORE COMPETENCIES

GenAI & LLMs: Production LLM Deployment , RAG Architecture , LLMOps ,vLLM, Guardrails & Safety , Prompt Engineering , Fine-tuning, Multi Agents, Quantization.
Leadership: Team Management , MLOps Strategy , AI Governance , Regulatory Compliance , Cross-functional Collaboration
ML Infrastructure: Real-time Pipelines , Model Monitoring , Drift Detection , Vector Databases , Semantic Search , CI/CD
Technical Stack: Python , PyTorch , TensorFlow , LangChain , Transformers , FastAPI , Docker , Kubernetes , Airflow
Cloud & Data Infrastructure: AWS, GCP , Azure, Typesense , Qdrant

PROFESSIONAL EXPERIENCE

Upstox | Mumbai

Senior Machine Learning Engineer | May 2022 - Present

- Lead 12-engineer ML team establishing MLOps best practices and fintech compliance frameworks aligned with financial industry standards and model risk management protocols.
- Architected AI Trading Copilot (https://upstox.com/contact-us/) serving 1M+ users, real-time news aggregation (sub-second latency), RAG + function calling powered system, and HITL escalation workflows.
- Deployed 20B parameter production LLM achieving 100 RPS throughput, **generating $108K annual cost savings** through optimized inference and infrastructure.
- Implemented comprehensive guardrails framework including PII detection, content filtering, prompt injection prevention, and jailbreak detection aligned with enterprise security standards.

- Built high-performance Typesense search infrastructure with <100ms P95 latency enabling semantic search and personalized recommendations for 1M+ users.
- Deployed Qdrant vector database for RAG chatbot with HNSW indexing and hybrid search capabilities
- Deployed Kubernetes-orchestrated Airflow managing 50+ DAGs with automated monitoring and drift detection, **reducing pipeline failures by 60%.**

Bajaj Allianz General Insurance | Pune

Data Scientist | Dec 2020 - Apr 2022

- Engineered intelligent document digitalization platform using AWS Textract OCR and custom NLP, **increasing operational efficiency by 40%** across 10K+ monthly documents with 92% accuracy
- Built quantitative investment analytics platform with portfolio optimization and forecasting for ₹500Cr+ assets under management
- Implemented knowledge graph-based fraud detection system for motor and health claims, **identifying 15% additional fraudulent cases** through relationship mapping
- Deployed real-time ML scoring models processing 1K+ daily claims with automated drift detection, **reducing investigation time by 45%**

Hopscotch Health | Mumbai

Data Scientist | Dec 2019 - Dec 2020

- Orchestrated healthcare data aggregation from 30+ hospitals using Airflow with automated data quality checks, establishing infrastructure contributing to **$50K revenue expansion**
- Developed CNN-based Indian food recognition model achieving 87% accuracy on 15K+ images for automated nutritional tracking
- Created marketing analytics dashboards with A/B testing framework optimizing campaign ROI by 25% through data-driven targeting

Moneycontrol / Network18 Media | Mumbai

Developer | Jun 2018 - Dec 2019

- Developed hybrid recommendation engine achieving 42% click-through rate using collaborative filtering and semantic search, **increasing engagement by 28%**
- Built NLP-based content moderation system filtering 85% of spam on Moneycontrol Forum with 94% precision

TECHNICAL SKILLS

ML/DL Frameworks: PyTorch, Transformers , BERT , XGBoost , LightGBM , Scikit-learn
GenAI & NLP: RAG , Langchain , Huggingface, RLHF , Semantic Search,Agents
MLOps & DevOps: Airflow , MLflow , FastAPI , Docker , Kubernetes , GitHub Actions
Databases: MySQL ,MongoDB , BigQuery , Athena/Presto , Typesense , Vector Databases

EDUCATION

Bachelor of Engineering, Information Technology | VIT, Mumbai | 2018

OPEN SOURCE & LEADERSHIP

Blazemetrics – Rust-powered LLM evaluation suite | https://blazemetrics.vercel.app/
MCPConn – MCP client library | https://github.com/2796gaurav/mcpconn
Contributor – Langchain, Pandas, Microsoft Qlib
Conference Speaker – AZConf: Regulatory Frameworks for LLMs in Fintech | https://2023.techxconf.com/
Technical Writer – Articles on Towards Data Science & Analytics Vidhya (500+ followers) | https://medium.com/@gauravc2708
Research – Predictive Analysis of Student Employability (IJSRCSEIT) | http://ijsrcseit.com/CSEIT1833228
YouTube Creator – ML paper explanations (200 subscribers) | https://www.youtube.com/@gauravchauhanml

CERTIFICATIONS

Data Engineering Nanodegree | Udacity