SHASHANK GUPTA

Bangalore, India • shashank5122@gmail.com • 9044652492 • linkedin.com/in/shashank-gupta-ai

PROFESSIONAL SUMMARY

Lead Machine Learning Engineer with 8 years of experience architecting and scaling production AI systems across cybersecurity, fintech, and enterprise platforms. Expertise in LLM-powered systems, RAG architectures, agent orchestration, and distributed model serving (vLLM, Triton). Delivered measurable business impact through 60% GPU cost reduction, production-grade observability, and robust MLOps practices. Strong track record of building reliable, cost-efficient, and high-performance AI systems from design to deployment.

PROFESSIONAL EXPERIENCE

HiLABS | Lead Data Scientist (LLM & AI Systems Engineering)

Bangalore | Mar 2025 – Present

- **Architected and deployed a production-grade LLM-powered Robocall Automation System** integrating LiveKit, Twilio, and backend microservices, orchestrated via **LangGraph-based multi-step agent workflows**. Implemented advanced prompt optimization and context engineering strategies, including structured system prompts, dynamic retrieval augmentation, function/tool calling workflows, and guardrail-based response control to reduce hallucinations and improve task reliability in production. Reduced manual verification workload by 60% while improving response accuracy and operational throughput.
- Designed scalable LLM serving infrastructure leveraging **vLLM and Triton Inference Server**, implementing NF4-quantized Qwen 2.5–7B models, batch-parallel inference, and context-window optimization to achieve 60% GPU cost reduction while maintaining latency and quality SLAs.
- Integrated **LiteLLM-based observability and cost monitoring**, enabling request-level tracing, token usage analytics, latency benchmarking, and model performance evaluation across staging and production environments.
- Built an enterprise-grade text standardization and entity resolution platform with intelligent caching, golden-record logic, and inference deduplication, cutting redundant API calls by 50% and significantly lowering compute costs.
- Established organization-wide MLOps foundations (MLflow, DVC, Docker, CI/CD), enabling reproducible pipelines, experiment tracking, and structured model lifecycle management across teams.

LONDON STOCK EXCHANGE GROUP (LSEG) | Senior Data Scientist

Bangalore | Feb 2023 – Mar 2025

- **Designed and led development of an enterprise-grade Retrieval-Augmented Generation (RAG) system** leveraging Llama 2, optimized embedding pipelines, and FAISS-based vector indexing to enable scalable internal knowledge retrieval. Integrated LangChain for orchestration and built conversational workflows with structured prompts and retrieval grounding, delivering high-quality contextual responses and winning first place at an internal AI innovation showcase.
- **Enhanced anomaly detection systems by evolving from traditional ML models (Isolation Forest) to deep generative approaches (Variational Autoencoders)**, improving detection performance by 15%. Implemented data drift monitoring and retraining pipelines to ensure production stability and model reliability over time.
- Built distributed data processing pipelines using **PySpark and Dask** to generate curated analytics heatmaps for content operations, reducing manual reporting effort by 60% and enabling faster data-driven decision-making.
- Mentored and guided junior data scientists, conducted design reviews, and led cross-functional planning sessions with business stakeholders to align AI solutions with operational and regulatory constraints.

FIREEYE TECHNOLOGY | Research Scientist

Bangalore | Sep 2021 – Feb 2023

- Architected and deployed machine learning–driven **Domain Generation Algorithm (DGA) detection systems** within enterprise security pipelines, improving threat classification accuracy and reducing false positives/negatives across large-scale DNS traffic environments.
- Designed end-to-end ML workflows spanning feature engineering, model experimentation (tree models, SVMs, RNNs), and production deployment, integrating detection systems into broader security monitoring frameworks.
- Built and deployed a real-time **DNS tunneling detection system on AWS SageMaker**, leveraging managed training jobs, scalable inference endpoints, and model versioning to enable low-latency threat detection at scale.
- Established structured **MLOps practices using MLflow in conjunction with SageMaker**, enabling experiment tracking, model artifact management, reproducible training pipelines, and controlled deployment workflows.

CLOUDSEK | Senior Machine Learning Engineer

Bangalore | Dec 2019 – Sep 2021

- Architected and deployed an automated large-scale data collection and threat intelligence platform integrating web crawling, NLP, and computer vision models, reducing manual processing effort by 90%.
- Optimized ML inference pipelines and containerized deployments (Docker), doubling system throughput without additional infrastructure and improving operational efficiency by 30%.
- Designed hybrid AI models combining CV, NLP, and structured ML techniques to accelerate threat detection by 25%, contributing to 15% client growth.
- Led and mentored a team of 4 ML engineers, conducted code reviews, and improved engineering productivity by 20%.
- Deployed scalable ML services on Google Cloud, integrating asynchronous messaging (RabbitMQ) and distributed processing for production-grade reliability.

BIGTHINX | Computer Vision Engineer

Bangalore | Jul 2018 – Jul 2019

- Developed computer vision–driven virtual try-on systems using pose estimation and semantic segmentation for realistic avatar-based fashion visualization.
- Built body measurement and outfit classification modules leveraging OpenPose, DeepLab segmentation, and CNN-based models.

EDUCATION

KRISHNA INSTITUTE OF ENGINEERING AND TECHNOLOGY (KIET) **Ghaziabad, UP**
Bachelor of Technology, Information Technology **2014-2018**

TECHNICAL SKILLS

- **LLM & Generative AI:** RAG architectures, LangGraph, Prompt Optimization, Context Engineering, Function/Tool Calling, Guardrails, PEFT (LoRA/QLoRA), Quantization (NF4), LiteLLM
- **Model Serving & Infrastructure:** vLLM, Triton Inference Server, Ray Serve, TensorFlow Serving, Docker, Kubernetes, AWS (SageMaker), Google Cloud
- **Vector & Data Systems:** FAISS, Vector Indexing & Retrieval Optimization, Embedding Pipelines, SQL, Snowflake
- **Machine Learning:** Anomaly Detection, Autoencoders, Tree-Based Models, NLP, Computer Vision, Time-Series, Transformers
- **Data & MLOps:** PySpark, Pandas, MLflow, DVC, CI/CD