

Vikash Agrawal

Address - Bangalore, India

Contact - agrawalvikash2015@gmail.com | +91 8871647511

LinkedIn- <https://www.linkedin.com/in/vikash-agrawal-027b14102/>

Github - <https://github.com/VikashAgrawal-DataScientist/>

Portfolio - <https://vikashagrawal.com/>

Summary

Senior AI Engineer with 7+ years of experience building production-scale AI systems, distributed platforms, and cloud-native applications, including 2+ years focused on Generative AI, Agentic AI, and RAG. Experienced in building multi-agent AI systems with planner-executor workflows, tool calling, memory (short-term and long-term), human-in-the-loop approval, and evaluation pipelines for enterprise use cases. Strong expertise in Python, AWS, LangChain, LangGraph, LangSmith, and scalable AI infrastructure, with hands-on experience deploying secure, observable, and cost-efficient LLM systems in regulated environments.

Experience

Senior AI Engineer – Solstice Intelligence

Jan 2026 – Present

- Designed and deployed an event-driven document processing platform on AWS using S3, Lambda, ECS Fargate, DynamoDB, Bedrock, and Docker, automatically processing 10K+ regulatory documents per month into structured AI-ready datasets.
- Architected and built an Agentic Regulatory Intelligence Platform for life-science clients, ingesting structured regulatory documents from agencies such as FDA and EMA, transforming large-scale multi-format documents into searchable knowledge assets using AWS-native event-driven pipelines, vector indexing, and metadata-aware retrieval to enable near real-time regulatory intelligence.
- Designed a multi-agent LangGraph-based reasoning workflow with specialized planner, retrieval, change-detection, impact-analysis, and report-generation agents to automatically detect semantic regulatory changes, localize section-level diffs, and generate explainable compliance impact assessments with citation-backed reasoning for downstream business.
- Productionized the AI platform on AWS using ECS Fargate, OpenSearch vector search, Bedrock LLMs, and LangSmith observability, implementing evaluation pipelines for retrieval quality, hallucination, latency monitoring, and cost optimization to deliver scalable low-latency regulatory insights.
- Provisioned and automated cloud infrastructure on Amazon Web Services using Terraform, managing ECS, Lambda, IAM, S3 event triggers, ECR, and DynamoDB, while building GitHub Actions–based CI/CD pipelines for infrastructure provisioning, Docker image builds, ECR publishing, and seamless multi-environment application deployments.
- Architected evaluation frameworks for LLM systems to quantify retrieval quality, factual correctness, and hallucination risk, enabling explainable and production-grade AI in regulated domains.

Senior Data Scientist / Generative AI Engineer– Scienaptic AI

Aug 2021 – Jan 2026

- Designed and deployed scalable RAG and semantic code intelligence platforms using LangChain, transformer embeddings, vector retrieval, and context engineering techniques, while establishing evaluation methodologies to measure retrieval relevance, answer quality, and system effectiveness at enterprise scale.
- Built and deployed large-scale credit risk models using Bureau and LexisNexis alternative data to predict probability of default, leading feature engineering, model validation, monitoring, and production deployment across lending portfolios.
- Analyzed millions of customer and credit records to identify risk patterns, behavioral signals, and

- emerging fraud indicators, improving portfolio risk management and decision-making.
- Developed anomaly detection and monitoring frameworks to identify model drift, data quality issues, and unusual customer behavior patterns.
- Incorporated fairness and bias mitigation techniques to align with regulatory compliance, including metrics like Adverse Impact Ratio (AIR) across gender, age, and race.
- Automated monthly swap reports and validation pipelines for models and strategies, enabling scalable monitoring across multiple versions.
- Owned end-to-end model pipeline:
 - o Data collection, EDA, feature engineering, feature selection, reject inference
 - o Hyperparameter tuning via Hyperopt/Optuna and model validation (KS, Gini)
 - o Model deployment, monitoring, and stability analysis using PSI and CSI
- Deployed containerized LLM inference services on AWS EC2 using Docker images, enabling consistent and scalable API endpoints for downstream applications.
- Defined and monitored AI quality metrics including retrieval precision, task success rate, hallucination frequency, response consistency, and user feedback signals to support continuous improvement initiatives.
- Evaluated multiple foundation models including Claude, GPT, and open-source LLMs for cost, latency, and answer quality optimization across enterprise AI workloads.

System Engineer – Tata Consultancy Services (TCS)

2019 – 2021

- Built backend services and REST APIs using Python (Django) and PostgreSQL.
- Developed full-stack applications using React.js and Python-based microservices.
- Participated in software development lifecycle including testing, deployment, and collaborative engineering workflows.

Skills

- Agentic AI - Planner-Executor Agents, Multi-Agent Orchestration, Tool Calling, Function Calling, Memory (STM/LTM), HITL, Reflection, Self-Correction, Guardrails, Agent Evaluation
- RAG - Data Ingestion, Hybrid Retrieval, Query Rewriting, Re-Ranking, Context Compression, Corrective RAG, Agentic RAG, Chunking, Embedding, Retrieval Evaluation
- AI Evaluation & Testing - LLM Evaluation, Benchmarking, Groundedness Evaluation, Hallucination Detection, Prompt Evaluation, Agent Testing, Quality Metrics Design
- AI Frameworks - Langchain, Langgraph, Langsmith
- Vector Store - Pinecone, Chroma, FAISS
- Machine Learning: Logistic Regression, Decision Tree, Random Forest, XGBoost, SHAP, PSI, CSI etc.
- Deep Learning: ANN, CNN, RNN, LSTM, GRU, Pytorch
- NLP: Text Preprocessing, Word2Vec, NER, TF-IDF, RAG, HuggingFace, Transformers, spaCy,
- Cloud: Amazon Bedrock, ECS, Lambda, EC2, SageMaker, S3, API Gateway, EventBridge, VPC, IAM, EBS, ELB

Certifications

- AWS Certified Machine Learning – Associate
- AWS Certified Cloud Practitioner
- Natural Language Processing Specialization – Coursera

Education

- B.E. in Computer Science | RGPV | 2019 | 7.2 CGPA
- 12th (MP Board) | 2015 | 83.2%
- 10th (MP Board) | 2013 | 85.6%

Achievements

- Data Science employee of the Year – 2023-24, Scienaptic AI
- State-Level Cricket Player (MP) & Baseball Player (MP)