KAUSHIK KUMAR KOLAR RAVINDRA KUMAR

+1 (862) 230-8954 | krkaushikkumar@gmail.com | linkedin.com/in/kaushikkumarkr | github.com/kaushikkumarkr | Harrison, NJ, USA

SUMMARY

AI Engineer & Product Strategist specializing in **Agentic AI, Compound AI Systems**, and **Scalable MLOps**. Expertise in architecting privacy-preserving RAG pipelines and deterministic multi-agent workflows that bridge the gap between advanced research and business outcomes.

PROFESSIONAL SKILLS

Languages: Python, SQL **Prompt Engineering:** Few-shot, Chain of Thought, ReAct, Reflection Patterns

Machine Learning & Deep Learning: PyTorch, Transformers, XGBoost, LightGBM, CNNs, LSTMs, Transfer Learning, Feature Engineering, SHAP, Optuna

Agentic AI: LangChain, LangGraph, Deepagents, CrewAI, Model Context Protocol (MCP), A2A, Multi-Agent Orchestration, n8n

RAG Systems: Hybrid Retrieval, Cross-Encoder Reranking, BM25, Semantic Search, Embeddings, Citation Validation, Chunking Strategies

Vector Databases: Qdrant, ChromaDB, pgvector, Pinecone, FAISS

DevOps: Docker, Kubernetes, MLflow, Pydantic, CI/CD, FastAPI, PostgreSQL

Observability: Arize Phoenix, Langfuse, Ragas, LangSmith, OpenTelemetry, LLM-as-a-Judge

Experimentation: A/B Testing, Hypothesis Testing, Causal Inference

AI Safety: Guardrails, Hallucination Detection, PII Filtering, Jailbreak Prevention, Confidence Scoring, Response Verification

Product & Project Management: Agile/Scrum, Roadmapping, PRD Writing, KPI Definition, GTM Strategy, User Research, JIRA, Cross-functional Leadership

WORK EXPERIENCE

FIELDWORKER.AI - AI Engineer | New Jersey, USA | *Remote* | *Internship* *Feb 2026 - Current*

- Architected an end-to-end agentic SDR parsing system using a **self-hosted open-source LLM** with vision and function-calling capabilities, **automating extraction** of 7+ structured entity types (customer, diagnosis, etc.) from DDD Service Detail Reports and reducing manual data entry time from hours to under 3 minutes per document.
- Designed and implemented agent-specific backend wrapper endpoints in Node.js/PostgreSQL that encapsulate check-then-create-or-update logic, enforcing strict **endpoint whitelisting and schema validation** to constrain agent operations within **defined boundaries, cutting unauthorized API access risk** by 100% and ensuring full audit trail coverage across all automated actions.
- Built a **human-in-the-loop (HITL) validation pipeline** integrating real-time entity verification, inline data correction UI, and sequential execution tracking across 8 workflow stages - maintaining data accuracy through **user-gated confirmation** while enabling async agent processing that **reduced frontend blocking time** and supported concurrent SDR workloads at scale.

CISCO SYSTEMS, INC. - Software Engineer Intern | Bangalore, Karnataka, India | *Onsite* | *Internship* *Feb 2024 - Jun 2024*

- Developed **scalable backend solutions** to enhance client business and financial operations, collaborating with the CCW Renewals team to deliver **3 new features** that streamlined processes and achieved a **20% increase in operational efficiency** across 500+ enterprise accounts.
- Designed and implemented **scalable microservices using Spring Boot**, incorporating **optimized caching mechanisms** and asynchronous processing to handle concurrent requests, resulting in a 30% improvement in response times during peak traffic, reducing database load by 45%.
- Refactored and optimized validation code by implementing comprehensive unit and integration tests using Mockito, introducing test-driven development practices leading to an **increase in code coverage from 73% to 95%,** reducing production errors by 40%, and establishing testing best practices adopted as team standards.

VERZEO EDUTECH PVT. LTD. - ML Engineer | Bangalore, Karnataka, India | *Remote* | *Internship* *Feb 2023 - Mar 2023*

- Developed **real-time AI models** using **CNN and RNN** for **image classification and time-series forecasting**, implementing data augmentation and hyperparameter tuning techniques that achieved an accuracy improvement of 15% in predictions while reducing model training time by 20%.
- Collaborated with cross-functional teams (product, engineering, data science) to deliver AI-driven solutions that improved decision-making processes and streamlined analytics workflows, **translated business requirements into ML pipelines** and presented model insights to non-technical stakeholders.
- Optimized **deployment of AI models on Google Cloud Platform (GCP)** using containerized microservices, implementing **batch inference and model caching** strategies that reduced latency by 25% and scaled resources effectively to support 50% higher workload capacity during peak usage.

EDUCATION

NEW JERSEY INSTITUTE OF TECHNOLOGY - Newark, NJ, USA *Sep 2024 - Present (Expected May 2026)*

Master of Science in Computer Science

B M S COLLEGE OF ENGINEERING - Bangalore, Karnataka, India *Sep 2020 - Jun 2024*

Bachelor of Engineering in Information Science and Engineering

PROJECTS

Antigravit | Enterprise Data Analysis Agent *Jun 2025 - Aug 2025*

- Built a multi-agent AI system using **LangGraph with 6 specialized nodes**, implementing intent-based orchestration with conditional branching and self-correcting SQL generation, achieving **100% routing accuracy, 83% SQL semantic equivalence, and 100% response faithfulness** across complex analytical queries.
- Engineered **Model Context Protocol (MCP) integration layer** with unified tool abstraction across heterogeneous sources (PostgreSQL, SQLite, Filesystem), featuring schema caching (60s TTL), connection pooling, and defense-in-depth security (SQL injection prevention, filesystem sandboxing, SELECT-only enforcement) with zero data exfiltration with 100% local execution.
- Designed end-to-end observability infrastructure using **Arize Phoenix with OpenTelemetry** distributed tracing, real-time WebSocket streaming, and a custom LLM evaluation framework (SQL Quality, Faithfulness via LLM-as-a-judge), enabling automated regression detection and continuous agent performance optimization.

RAG Foundry | Enterprise-Grade Agentic RAG Framework *Jan 2025 - Mar 2025*

- Developed a **7-stage Agentic RAG pipeline** integrating Hybrid Retrieval (Qdrant + BM25), Cross-Encoder reranking, and **4 MCP connectors** (Drive, Notion, Obsidian, Filesystem) for autonomous multi-platform retrieval, achieving **0.98 Relevancy and 1.00 Faithfulness** on Ragas evaluation.
- Designed an end-to-end LLM Safety System featuring **6 custom guardrails** (Jailbreak, PII, toxicity, refusal, hallucination filtering) and a **Citation Validation pipeline** that automatically extracts references and flags phantom citations, effectively mitigating hallucination risks across generated responses.
- Implemented a **5-signal weighted confidence scorer** and **Langfuse observability** to capture 15+ real-time metrics across 7 nested spans, while deploying optimized Qwen2.5-7B-4bit inference on Apple Silicon (MLX), validated by a 100% unit test pass rate.