

Akshay Katageri

Generative AI Engineer specializing in Multi-Agent Systems and RAG

+1 (864)-837-0224 | akshaykatageri@gmail.com | Jersey City, NJ | [LinkedIn](#) | [Portfolio](#)

CAREER Highlights

4 years of expertise in Advanced AI concepts delivering production-ready solutions resulting in reduced cost, speed delivery, and boost in accuracy. Recognized for building safe, observable AI workflows that deliver measurable ROI, earn executive & stakeholder trust with end-user adoption. Proven cross-functional collaborator, translating deep technical skills into business impact that resonates with executives, clients & stakeholders.

TECHNICAL SKILLS

- **Agent Architecture:** LangGraph, CrewAI, AutoGen, OpenAI Assistants & Agents SDK, Google ADK.
- **Protocols & Orchestration:** MCP servers/clients, Multi-Agent Systems, Multimodal AI.
- **Agentic Patterns:** ReAct, Self-Correction Patterns, Tool/Function Integration, Automated Multi-step Planning.
- **Data & Vector DB:** Hugging Face, FAISS/Milvus/Pinecone/Chroma/Astra.
- **Prompting & Alignment:** Prompt/Context Engineering, Fine/Instruction Tuning Techniques.
- **RAG & Memory:** Haystack, Hybrid Retrieval, RRF Rerankers, LlamalIndex Pipelines & Memory, Multi-hop QA.
- **Evaluation & Observability:** LangSmith/Langfuse, Arize Phoenix, RAGAS/TruLens, Prompt Injection, Adversarial Testing.
- **Serving & Performance:** vLLM, TensorRT- LLM/Triton, ONNX Runtime/TGI, FlashAttention-2, KV-Cache, Speculative Decoding.
- **Safety:** Prompt & PII Filtering/Injection Defense Tools, Audit Trails, Global Regulation Compliance.
- **Systems & MLOps:** Scikit-learn, PyTorch, Tensorflow, Keras, spaCy/NLTK, Docker/Kubernetes, CI/CD, AI/ML.
- **Cloud DevOps:** AWS (Sagemaker, Bedrock)/GCP/Azure pipelines.
- **Soft Skills:** Critical Thinking & Problem Decomposition, Rapid Prototyping & Experimentation, Cross-Functional Collaboration, Context Translation, Adaptability & Strategic Agility, Ethical Judgment & Critical Reflection, Model Version Control, Monitoring & Resilience.

WORK EXPERIENCES

KGS Technology Group, Inc. (GenAI Engineer)

July 2025 - Present

- Redesigned Multi-Agent Workflows for client software using LangGraph & TensorRT; cutting P95 latency 45%, doubling throughput, and reducing task cost 33% at production scale.
- Slashed integration lead time by 90% (3 weeks to 3 days) by developing secure MCP servers with CrewAI orchestration.
- Achieved zero PII incidents across 18+ tool calls by implementing rigorous audit trails and PII Filtering Defense Tools.
- Launched Hybrid Retrieval RAG, improving groundedness +12 pts, accuracy +9 pts, and deflecting 22% of tickets, saving \$3.8k/month in OpEx.

CloudData Technology (Data Analyst)

August 2024 - July 2025

- Created LLM-driven pipelines with LangGraph and LangSmith tracing, reducing forecast error 18%, false positives 27%, and boosting precision +11 pts.
- Shipped NL→SQL analytics copilot automating Power BI workflows, eliminating 12 hrs/week manual ELT, cutting time-to-insight 55%, and increasing adoption +35%.
- Implemented AI observability & LlamalIndex RAG assistant, improving data integrity +23% and cutting incidents 60% while maintaining SLA 99.7%.

Bharat Soft Solutions Pvt. Ltd. India (Jr. Data Scientist)

Aug 2021 - April 2022

- Boosted F1 score 0.71→0.83 and reduced inference latency 35% using TensorFlow pipelines, transfer learning, augmentation, and quantization.
- Scaled dataset 5x and reduced experiment cycle time 7→3 days with reproducible MLflow workflows and parameterized training scripts.
- Packaged models as Dockerized Flask microservices, reducing analyst review time 30% across 3 variants.

Connection Loops Pvt. Ltd. India (Jr. Data Scientist)

Sept 2020 - June 2021

- Owned SQL pipelines with validation checks, reducing prep time 58% and stabilizing refresh reliability to 99.4% for flagship product.
- Built scikit-learn baselines with targeted feature engineering, improving AUC 0.74→0.82 and enabling earlier case detection.
- Delivered KPI dashboards for clinical stakeholders, shortening decision lead-time by 3 days and enabling 3 new research cohorts.

PROJECTS

Trip Planner – Multimodal, Multi-Agent Travel Copilot:

- Built a travel copilot using CrewAI & LangGraph boosting user satisfaction to 97% across 1,200+ itineraries.
- Improved trip planning speed by 4x with custom retrieval Haystack/LlamalIndex, cutting average time to 10 minutes.
- Optimized serving using vLLM, KV-cache, and Speculative Decoding, achieving 3.2s latency, 97% visa accuracy, and 22% cheaper itineraries.

Autonomous Delivery Agent – Route-Aware Ordering & Dispatch:

- Deployed Autonomous Delivery Agents with MCP protocols and vLLM, increasing route efficiency by 48% for 5k+ monthly orders.
- Quantized ONNX/TGI models, lowering compute costs by 35%, improving order success by 9% & 0 PII incidents across 12+ tool calls.
- Built-in automated audit and safety checks, reducing simulated order errors by 72% and demonstrating strong compliance readiness.

Universal Commerce Copilot – Chat-First Recommender & Checkout:

- Developed with OpenAI SDK & Vector search, lifting test checkouts by 3.2x while handling pricing, coupons and account workflows.
- Raised recommendation accuracy to 72% by hybrid retrieval RAG, also achieving 98.5% price/stock accuracy, P95 latency 2.8s & automated checkout saving \$18/order.

Education

Pace University, New York City, USA: Master's in Data Science

Savitribai Phule Pune University, Pune, India: Bachelor's in Computer Science