Akshay Katageri

AI Engineer specializing in Multi-Agent Systems and RAG

+1 (864)-837-0224| akshayykatageri@gmail.com | Jersey City, NJ |LinkedIn | Portfolio

CAREER Highlights

AI Engineer & Data Scientist with over 3.5 years of experience designing scalable, production-grade AI/ML workflows across healthcare and enterprise domains. Adept at reducing latency, integrating secure systems and driving adoption of advanced AI solutions, including multi-agent orchestration, hybrid retrieval and analytics automation. Strong collaborator and mentor, known for delivering clear business value such as slashing integration time, improving prediction accuracy and enabling faster, safer decision-making while fostering a culture of technical growth and teamwork. Comfortable partnering across engineering, product and compliance to ensure solutions meet reliability and safety standards for real-world impact.

TECHNICAL SKILLS

- **Agent Architecture:** LangGraph, CrewAI, OpenAI Assistants & Agents SDK.
- **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/Pinecone/Chroma/Astra.
- **Prompting & Alignment**: Prompt/Context Engineering, Fine/Instruction Tuning Techniques.
- **RAG & Memory**: Haystack, Hybrid Retrieval, RRF Rerankers, LlamaIndex Pipelines & Memory.
- **Evaluation & Observability**: LangSmith/Langfuse, Arize Phoenix, Prompt Injection.
- **Serving & Performance**: vLLM, TensorRT- LLM/Triton, ONNX Runtime/TGI.
- **Safety**: Prompt & PII Filtering/Injection Defense Tools, Audit Trails, Global Regulation Compliance.
- **Systems & MLOps**: Scikit-learn, PyTorch, Tensorflow, Keras, Docker/Kubernetes, AI/ML.
- **Soft Skills**: Problem Decomposition, Rapid Prototyping, Experimentation, Cross-Functional Collaboration, Ethical AI Judgment

WORK EXPERIENCES

Elevance Health – USA (AI Engineer) **July 2025 - Present**

- Designed and optimized multi-agent workflows using LangGraph/TensorRT-LLM, reducing P95 latency by 45% & doubling throughput at scale.
- Developed secure MCP-based orchestration that cut system integration time from 3 weeks to 3 days, improving internal adoption.
- Embedded strict PII filtering and audit trails, enabling 18+ tool integrations with zero safety violations across production workloads.
- Built a Hybrid Retrieval RAG system that improved groundedness by +12 points and reduced support escalations by 22%, translating to ~$3.8k/month OpEx savings.
- Partnered with product and compliance teams to align agentic decisions with safety, reliability and clinical accuracy standards.
- Developed a Generative AI-powered chatbot using OpenAI's API and LangGraph that handles 500+ weekly client service queries automatically, improving response time by 65%.

Mphasis – India (Data Scientist) **August 2024 - July 2025**

- Built predictive pipelines using scikit-learn, pytorch and tensorflow, improving F1 scores 0.71->0.83 with transfer learning and model optimization.
- Reduced inference latency 35% through quantization and ONNX optimizations, enabling faster client-facing applications.
- Created reproducible MLflow workflows that cut experimentation cycle time from 7 days to 3 days, enabling faster iteration.
- Delivered KPI dashboards that improved stakeholder decision-making speed by 3 days and supported new clinical insights.
- Implemented LLM-powered analytics pipelines with LangGraph and LangSmith, reducing forecasting error 18%, cutting false positives 27% and improving precision across automated workflows.
- Created an NL→SQL analytics copilot that automated Power BI/ELT tasks, eliminating 12 hours per week of manual reporting work and improving time-to-insight by 55%.
- Increased analytics tool adoption by 35% across teams by delivering intuitive AI-assisted insights and reliable workflow automation.

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/LlamaIndex, 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.

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

Certifications

- Advanced LLMs with Retrieval-Augmented Generation (RAG)
- Agentic AI Design Patterns for Generative & Predictive AI
- Time Series Analysis & Forecasting with GPT-4
- Synthetic Data for AI Privacy, Fairness & Explainability