Pratham Rana

AI Engineer | Databricks Certified Generative AI Engineer Associate

Noida, India | +91 85951 98755 | ranapratham213@gmail.com

linkedin.com/in/pratham-rana-a512b2202 | github.com/ranagh401 | prathamrana.vercel.app

PROFESSIONAL SUMMARY

AI Engineer with 2.5 years of experience building enterprise Generative AI applications, multi-agent platforms (LangGraph, MCP), RAG and NL-to-SQL conversational analytics, and decision-intelligence systems on Azure and AWS. Delivered production LLM platforms for healthcare and retail clients supporting 500+ business users.

TECHNICAL SKILLS

Programming & Backend: Python, SQL, FastAPI, SQLAlchemy, asyncio, Pydantic, Uvicorn, REST APIs, Streaming APIs, WebSockets, Git, pytest, Unit Testing

Generative AI & LLMs: Large Language Models (LLMs), Azure OpenAI, OpenAI API, Anthropic Claude API, Google Gemini, Hugging Face Transformers, PyTorch, LangChain, LangGraph, LlamaIndex, Prompt Engineering, Context Engineering, Structured Outputs, Function Calling / Tool Calling, Fine-Tuning (LoRA, PEFT), Multimodal LLMs

Agentic AI: AI Agents, Multi-Agent Systems, Agent Orchestration, Model Context Protocol (MCP), MCP Servers & Clients, Agent-to-Agent (A2A) Protocol, OpenAI Agents SDK, CrewAI, AutoGen, Agent Memory, Planning & Reasoning, Tool Binding, Sandboxed Tool Execution, Voice Agents (STT/TTS), Agent FinOps (Token Metering, Cost Attribution, Spend Forecasting, Budgets), Conditional Routing, Checkpointing, Human-in-the-Loop, Guardrails, SSE, Socket.IO

RAG & Retrieval: Retrieval-Augmented Generation (RAG), Agentic RAG, Embeddings, SentenceTransformers, Vector Search, Hybrid Search, Semantic Search, Reranking, Chunking Strategies, Knowledge Graphs, Semantic Caching

LLMOps & Evaluation: LLM Evaluation, LLM-as-a-Judge, RAGAS, LangSmith, Langfuse, Observability & Tracing, Prompt Versioning, A/B Testing, Token & Cost Optimization, Latency Optimization, Responsible AI, Guardrails, PII Detection & Redaction, Hallucination Mitigation

Cloud & MLOps: Azure AI Foundry, Azure AI Search, Azure Key Vault, Azure Blob Storage, Azure SQL, AWS Bedrock, AWS S3, IAM, KMS, Secrets Manager, Docker, Kubernetes, Azure Kubernetes Service (AKS), CI/CD, GitHub Actions, Linux, MLOps

Databases & Vector Stores: PostgreSQL, Snowflake, Databricks SQL, Azure SQL, PgVector, Qdrant, Pinecone, FAISS, Chroma, Redis

Enterprise Integrations: Microsoft Teams Bot Framework, Adaptive Cards, Microsoft Graph API, OAuth2, MSAL, JWT, Role-Based Access Control, Microsoft Copilot Studio

PROFESSIONAL EXPERIENCE

AI Engineer | Polestar Analytics

Noida, India | October 2024 – Present

- Designed and delivered enterprise AI platforms for healthcare and retail clients, translating business requirements into scalable LLM-powered applications and analytics workflows used by 500+ business users.
- Engineered agentic systems with LangGraph, LangChain, FastAPI, and enterprise LLM providers, including MCP-based tool execution, sandboxed tools, guardrails, PII detection, agent FinOps, voice interaction, conditional routing, state management, and observability; deployed on Docker and Azure Kubernetes Service (AKS).
- Built conversational analytics and NL-to-SQL capabilities over enterprise data platforms, enabling business users to explore data, ask contextual follow-up questions, and generate actionable insights.

KEY PROJECTS

Agenthood – Multi-Agent Workflow Orchestration Platform

Polestar Analytics

- Architected the backend orchestration layer with FastAPI and LangGraph, defining graph-based execution flows for sequential and parallel agent tasks, conditional branching, and state transitions.
- Integrated Model Context Protocol (MCP) servers as a standardized tool layer, enabling agents to discover and invoke enterprise tools, data sources, and APIs through MCP clients with function calling.
- Implemented agent guardrails and PII detection/redaction on inputs, tool calls, and outputs, enforcing content policies and data-privacy controls across workflows.
- Built FinOps for agents: per-agent token metering, cost attribution by model and provider, budgeted spend with alerts, and dashboards of actual vs. forecasted LLM cost per agent.
- Enabled voice-based interaction with agents through speech-to-text and text-to-speech pipelines, supporting real-time voice chat alongside text channels.

AI Insight Portal – Decision Intelligence Platform

Retail Client

- Developed configurable signal frameworks for KPI monitoring, threshold evaluation, severity scoring, and signal execution across enterprise retail datasets.
- Implemented automated anomaly and variance analysis workflows to detect unusual KPI movements and trigger downstream insight generation.

Healthcare Conversational Intelligence Platform

Biopharmaceutical Client

- Built LLM-driven conversational workflows with LangChain and FastAPI to translate user questions into contextual analytical requests over enterprise healthcare data in Snowflake.
- Achieved sub-2-second P95 latency for real-time analytical workloads under high concurrency; reduced the analytics backlog by 40% and accelerated self-service insights for business stakeholders.

EDUCATION & CERTIFICATION

Databricks Certified Generative AI Engineer Associate — Issued Jul 2026 · Valid to Jul 2028 · ID 187987580

B.Tech, Information Technology | Maharaja Surajmal Institute of Technology — 2020 – 2024 | 91.2%