Madan Raj S

Backend & AI Engineer — GenAI — LLM Systems & Infra — MLOps
madanrajs.pro@gmail.com — linkedin.com/in/0xmadan-raj — github.com/maximus0x0

Professional Summary

AI/Backend Engineer with 4 years of experience building production-grade, distributed AI systems, including multi-agent LLM architectures, RAG pipelines, and event-driven microservices, at Mercedes-Benz R&D. Strong background in asynchronous Python (FastAPI, Pydantic), agent orchestration (LangChain, LangGraph, MCP JSON-RPC), event-driven messaging (Kafka), and containerized deployment (Docker, Kubernetes). Proven track record of shipping enterprise-scale AI platforms serving 10K+ monthly queries, with strong foundations in data engineering, model evaluation, and MLOps.

Experience

Mercedes-Benz Research & Development India *Bangalore, India*
Senior AI Engineer *Apr 2026 – Present*

- Built a knowledge graph modelling companion agent that parses technical documentation, maps entities and relationships to standard ontologies, and recommends or applies graph model changes with human-in-the-loop verification; manual edits trigger automated impact analysis and propagation suggestions.
- Developed an agentic workflow for automated Workday test case generation, processing structured and unstructured data sources to produce comprehensive validation scenarios.

Mercedes-Benz Research & Development India *Bangalore, India*
AI Engineer / Backend Engineer *Apr 2023 – Mar 2026*

- Designed modular, agent-based multi-agent AI architecture using LangChain, MCP (JSON-RPC based orchestration), and a Neo4j knowledge graph, enabling specialized agent workers for intent-driven routing, knowledge retrieval, and multi-hop reasoning over structured and unstructured enterprise data.
- Engineered data preprocessing and retrieval pipelines for 500+ documents powering a production RAG platform – including chunking, embeddings, metadata enrichment, deduplication, and vector indexing – serving 10K+ monthly queries at 92% grounded response accuracy.
- Built and maintained containerized Python microservices (Flask, PostgreSQL, Docker, Kubernetes) supporting internal data-driven and AI-powered applications, with readiness/liveness health checks for reliable orchestration.
- Designed high-volume, event-driven data ingestion pipelines using PySpark, Databricks, dbt, and Airflow DAGs, processing 200k+ records daily with scheduling, dependency management, and automated failure recovery.
- Designed event-driven pipelines using Kafka pub/sub for asynchronous, reliable communication across 20+ microservices, reducing coupling between ingestion, processing, and downstream consumers.
- Developed automated data quality checks using rule-based and LLM-based validation (LLM-as-a-Judge), detecting anomalies, schema drift, and inconsistencies before production deployment.
- Built API-driven data ingestion systems and internal REST connectors, performing schema normalization, transformation, and request validation for downstream analytics and AI systems.
- Implemented data enrichment pipelines combining ML models and LLMs (Azure OpenAI, ChatOpenAI-style integrations) to enhance raw datasets with semantic signals and predictions, applying prompt engineering and prompt-template design.
- Engineered backend observability services for a data platform in Golang – centralized logging, metrics collection, and health monitoring across distributed pipelines – reducing incident response time.
- Established CI/CD pipelines automating build, test, and deployment for data and AI workflows using GitHub Actions, Docker, and Kubernetes.
- Implemented API authentication using OAuth2 flows for internal service-to-service and client-facing endpoints, following secure coding practices for credential handling.
- Instrumented AI/backend services with OpenTelemetry for distributed tracing and Prometheus for metrics collection, enabling proactive monitoring of pipeline and service health.

- Used LangSmith to trace, debug, and evaluate LangChain agent workflows, improving visibility into multi-agent execution paths and prompt performance.
- Built custom automation workflows using n8n to orchestrate data movement and reduce manual operational overhead across teams.
- Built a Databricks migration agent converting legacy ODI SQL to dbt-compatible queries at ˜90% accuracy, with automated data validation (row counts, schema checks, hashed comparisons).

Mercedes-Benz Research & Development India *Bangalore, India*
Graduate Engineering Trainee *Aug 2022 – Mar 2023*

- Developed PySpark-based distributed data pipelines for ML workflows, including feature engineering and large-scale data processing.
- Built predictive models (XGBoost, Gradient Boosting) on structured datasets, with strong focus on data preprocessing and feature quality.
- Implemented centralized logging, monitoring, and observability pipelines using Kafka, Elasticsearch, and Kibana across 20+ microservices.

Mercedes-Benz Research & Development India *Bangalore, India*
Engineering Intern *Mar 2022 – Jul 2022*

- Developed RESTful APIs and data access layers using Flask and PostgreSQL with containerized deployments.
- Built enterprise integration workflows handling structured data exchange using Oracle SOA Suite.

Projects

Agentic RAG Research Assistant *LangGraph, GPT-4, Pinecone, Tavily API*

- Designed an autonomous research assistant orchestrating web search, document retrieval, and code execution via a LangGraph-based agent state machine with self-refinement loops.
- Built a hybrid RAG pipeline (semantic + keyword retrieval) over 10K+ documents with improved grounding; developed an evaluation framework using RAGAS metrics.

Multimodal LLM Application – Visual QA *GPT-4o, OpenCLIP, Tesseract OCR, ChromaDB, FastAPI*

- Built a multimodal system answering queries over images, PDFs, and text using async FastAPI endpoints.
- Implemented OpenCLIP-based semantic image retrieval with OCR pipelines, and a real-time streaming interface (SSE-based token streaming) with conversation memory.

Fine-Tuned Domain-Specific LLM *Phi-3.5 Mini, LoRA, MLX, W&B*

- Fine-tuned Phi-3.5 Mini using LoRA on 10K+ instruction samples; tracked experiments using Weights & Biases.
- Optimized inference latency with KV-cache reuse and lazy loading.

Real-Time Fraud Detection System *PyTorch, Kafka, Redis, Kubernetes*

- Built a real-time, event-driven fraud detection system using LSTM autoencoders and Isolation Forests on a Kafka streaming pipeline.
- Designed a low-latency, horizontally scalable ML pipeline with Redis caching and Kubernetes autoscaling.

Education

Nitte Meenakshi Institute of Technology, Bangalore, India Jun 2018 – Aug 2022
Bachelor of Engineering in Computer Science – GPA: 8.4/10

Technical Skills

Languages: Python (async/await, type hints), Go, SQL, Bash
Backend: FastAPI, Flask, REST API design, AsyncIO, Pydantic (request validation), Microservices
GenAI: LangChain, LangGraph, LlamaIndex, MCP (JSON-RPC based orchestration), RAG, Prompt Engineering, Agentic AI / Multi-agent workflows, Knowledge Graphs, Intent Classification & Routing

LLMs: OpenAI, Azure OpenAI, Anthropic Claude, Llama, Mistral, Gemini
Retrieval: Vector Search, Hybrid Search, Semantic Search, Embeddings, Retrieval Pipelines
Messaging & Streaming: Kafka (pub/sub, event-driven pipelines), Server-Sent Events (SSE), Streaming responses
Infrastructure: Docker, Kubernetes (health probes, ConfigMaps/Secrets), GitHub Actions, Redis, PostgreSQL, Neo4j, MongoDB, Elasticsearch
Data: PySpark, Databricks, dbt, Airflow, Delta Lake
Cloud: Azure, AWS – cloud-native, containerized deployment
MLOps: MLflow, W&B, DVC, Docker, Kubernetes, CI/CD
Observability: OpenTelemetry (distributed tracing), Prometheus (metrics collection), LangSmith (LLM/agent tracing), Centralized logging, application & health monitoring (Kafka/Elasticsearch/Kibana)
Security: OAuth2, API Key Auth, secure credential handling, input validation
Developer Tools: Cursor, Claude Code, GitHub Copilot, Windsurf, Devin, Ollama

Certifications

Microsoft Azure AZ-900, DP-900 — Machine Learning Specialization – Coursera

Awards

Bronze Star @ Mercedes-Benz – Contributions in AI/ML and GenAI-based POCs, delivering solutions for the HR DataFactory workstream.