

BALAKUMARAN R

GenAI / AI Engineer | LLM Applications • Agentic AI • RAG • LLM Observability

Chennai, India | Open to Relocation | balakumaranrinfo@gmail.com | +91 96772 66036 | linkedin.com/in/bala0423 | balakumaran.in

PROFESSIONAL SUMMARY

GenAI / AI Engineer with 8 years of experience, currently building production LLM and agentic AI systems at Johnson Electric. Delivered a PLM support chatbot rolled out to 3,000 users (~500 active) — evolving it from vector-based RAG to LLM-driven retrieval to improve accuracy — and an MCP-based AI agent on Azure OpenAI GPT-5 with a Neo4j knowledge graph, serving ~600 engineering and quality users, with LLM observability through Arize Phoenix and Grafana. Combines hands-on experience in LangGraph, FastAPI and Azure AI Foundry with production ML on time-series data and a deep data engineering background (Databricks, PySpark, Kafka, Airflow). Mentors junior engineers and partners directly with business stakeholders.

CORE SKILLS

LLMs & Platforms: Azure OpenAI (GPT-5, GPT-4, GPT-4o mini), Azure AI Foundry, OpenAI API

GenAI Engineering: RAG, LLM-based retrieval, chunking, Agentic AI, LangGraph, LangChain, Model Context Protocol (MCP), FastMCP, Tiny Agents

Knowledge & Vector Stores: Neo4j (knowledge graph), Marqo, ChromaDB

LLMOps & Observability: Arize Phoenix (tracing), Grafana (usage, token cost), Power BI usage analytics

Apps & APIs: Python, FastAPI, Flask, REST APIs, Open WebUI, Chainlit, Streamlit, Microsoft Graph API

Machine Learning: Time-series modeling, LSTM, Kalman filter, NLP, Rasa NLU, TensorFlow, scikit-learn, Pandas, NumPy

Data & Cloud: Azure Databricks, PySpark, Delta Lake, Kafka, Airflow, Azure Data Factory, Snowflake, AWS Glue, BigQuery, SQL, MongoDB, PostgreSQL, Azure, AWS, GCP, Docker, Jenkins, CI/CD

PROFESSIONAL EXPERIENCE

Data Science Engineer | Johnson Electric

Jan 2024 – Present

Chennai, India (Hybrid)

- Mentor a team of 3 junior engineers; present solutions to stakeholders and work directly with business users.

Product Knowledge AI Agent (Agentic AI, Production) — GPT-5, Neo4j, FastMCP, Tiny Agents, Arize Phoenix, Grafana

- Built and deployed an agentic chatbot on Azure OpenAI GPT-5, giving ~600 users (500 engineering, 100 quality) self-service access to product information on motors and actuators.
- Modeled product data in a Neo4j graph database and exposed agent functions as FastMCP (Model Context Protocol) tools, orchestrated with Tiny Agents.
- Implemented LLM observability with Arize Phoenix to trace agent tool usage and responses; built Grafana dashboards to monitor usage, token cost and token expiry.

PLM Support Chatbot (GenAI, Production) — Azure OpenAI, Marqo, LangGraph, FastAPI, Open WebUI, Power BI

- Built an end-to-end GenAI support chatbot for PLM users, extracting SharePoint pages and wiki documentation via Microsoft Graph API; rolled out to 3,000 users with ~500 active users.
- Developed the initial RAG pipeline with Marqo as the vector database and Azure OpenAI, orchestrated with LangGraph and served through FastAPI.
- Redesigned retrieval to improve answer accuracy, replacing vector search with LLM-based chunk selection: GPT-4o mini selects relevant JSON chunks per query and a GPT-4 agent generates the final response, delivered through Open WebUI.
- Integrated in-chat ticket creation to escalate unresolved queries; validated bot responses with the L1 support team, resulting in fewer support tickets raised.
- Built a Power BI dashboard to analyze chatbot usage, user activity and query trends.

Predictive Models on Time-Series Test Data — Kalman filter, LSTM

- Water pump defect prediction (production): deployed a Kalman filter-based predictive model in production to detect water pump defects from real-time time-series test data, enabling earlier identification of defective units and reducing overall test time.
- Actuator endurance failure prediction (proof of concept): built an LSTM model to predict the cycle count at which actuators fail, using time-series log data from endurance testing, enabling earlier failure detection.

Career Break | Health and well-being

Aug 2023 – Jan 2024

Software Engineer | Resulticks

May 2023 – Aug 2023

Chennai, India

- Built ETL pipelines on Google Cloud Platform, loading and processing data in BigQuery.

Data Engineer | bitsCrunch (Germany)**Jun 2022 – Apr 2023***Remote from India*

- Built real-time Kafka streaming pipelines ingesting multi-chain blockchain data (NFTs, digital assets) into Azure Blob Storage; preprocessed data with PySpark on Azure Databricks.
- Maintained Airflow pipelines supporting the company's multi-chain NFT analytics platform.

OptiSol Business Solutions**Aug 2019 – Jun 2022***Chennai, India***Data Engineer***Sep 2021 – Jun 2022*

- Nutrien (Australia): Built Spark pipelines on Azure Databricks migrating enterprise data from SAP, Plasma, FMIS, Rewards, PostgreSQL and AWS S3 into Delta Lake tables; built data dictionaries.

Machine Learning Engineer*Aug 2019 – Sep 2021*

- E-Watercooler (USA): Built a Rasa NLU conversational AI chatbot in Microsoft Teams (intent classification, entity extraction, dialogue flows) that used Microsoft Graph API to schedule group calls on shared interests; deployed via Docker and Jenkins CI/CD with Azure Key Vault.
- MachineVantage (USA): Developed Flask RESTful APIs on AWS with Jenkins deployment, managed ~1 TB of data in MongoDB, built Airflow pipelines, and set up API monitoring with root cause analysis.
- MachineVantage (USA): Built Azure Data Factory ETL pipelines combining FMCG sales, Google Ads, demographic and census data into OLAP cubes analyzed in Power BI.

Software Engineer | Trumatics Software Labs**Jun 2018 – Jul 2019***Chennai, India*

- Built a decision tree price prediction model in Python served via Flask API on Linux and Heroku; automated API testing with Selenium.

EDUCATION

M.E., Systems Engineering and Operations Research — 79%**2018**

College of Engineering, Guindy (CEG), Anna University, Chennai

B.E., Computer Science and Engineering — 80%**2015**

Meenakshi Academy of Higher Education and Research, Chennai