

Ritesh Zambre

Pune, India | riteshzambre111@gmail.com | +91 70282 24490 | [linkedin](#) | [github](#) | riteshzambre.com

Summary

Senior AI/ML Engineer with 6 years of experience designing and shipping production AI systems on Microsoft Azure for regulated industries. Depth in Generative AI system design RAG architecture, retrieval and evaluation pipelines, agentic workflows alongside end-to-end MLOps and Microsoft Fabric data engineering. Delivered AI and analytics platforms for the World Health Organization under healthcare compliance constraints, plus production ML for credit risk and demand forecasting. Builds AI systems that are governed, observable, evaluated and cost-controlled.

Experience

- ❖ **Software Engineer (Senior AI/ML Engineer), e-Zest Solutions** – Pune, India Sept 2023 – Present
 - Architected a production Retrieval-Augmented Generation platform on **Azure OpenAI, Azure AI Foundry and Azure AI Search**, covering document chunking strategy, embedding generation, hybrid (vector + keyword) retrieval and semantic ranking to enable natural-language querying over enterprise knowledge repositories.
 - Owned retrieval and response quality end to end chunk-size and overlap tuning, metadata filtering, query rewriting, retrieval-parameter optimization and citation-based grounding measurably reducing ungrounded and hallucinated responses in evaluation runs.
 - Built an **AI-powered GitHub Pull Request Assistant** that retrieves contextual repository knowledge and generates code-review recommendations, combining RAG with tool-use orchestration over Azure OpenAI.
 - Designed and implemented enterprise **Azure MLOps pipelines** (Azure Machine Learning, Azure DevOps, GitHub Actions, MLflow) automating training, model registry promotion, deployment and monitoring, reducing deployment time by **50%**.
 - Designed the **security architecture** for enterprise AI workloads using Managed Identity, Service Principals, Azure Key Vault and scoped RBAC, meeting healthcare data-protection and compliance requirements.
 - Engineered distributed healthcare data pipelines with **Azure Databricks (PySpark)**, Azure Data Factory and **Microsoft Fabric Lakehouse** using medallion architecture, processing large-scale structured and semi-structured datasets for downstream AI workloads.
 - Delivered enterprise AI and analytics solutions for the **World Health Organization (WHO)** supporting global public health programmes across COVID-19, Tuberculosis (TB) and Antimicrobial Resistance (AMR), Neglected Tropical Diseases (NTDs), Malaria, Influenza.
 - Partnered with **Microsoft engineers** to migrate WHO Power BI workspaces to **Microsoft Fabric**, re-platforming onto Lakehouse architecture, Fabric Pipelines, Direct Lake semantic models and capacity-aware workload design.
 - Mentored engineers on Azure AI, Microsoft Fabric, MLOps and enterprise Generative AI, and authored reusable AI engineering standards and reference architectures adopted across teams.
- ❖ **Senior Data Scientist, Voiceback Analytics** – Pune, India Sept 2021 – Sept 2023
 - Led end-to-end data engineering and forecasting for **Theobroma (Food Retail)** built Azure Data Factory ETL pipelines for inventory, sales and stockout data, Power BI dashboards and SKU-level demand forecasting models validated by walk-forward backtesting. Improved forecast accuracy by **20%**, cut stockouts by **7%** and influenced **16 Cr. INR** in revenue.
 - Deployed **3 production credit-risk models** (Python, Azure ML) for a **microfinance lender** predicting loan default and forecasting 12-month expected credit loss (ECL) handled class imbalance, probability calibration and threshold selection, and used SHAP for model explainability and business sign-off. Contributed to a **0.36% reduction in credit losses**.
 - Designed a rule-based hourly production planning engine (Python) for **Paradise Food Court**, served via Azure Functions with a Power Apps input layer and Power BI reporting, reducing food wastage by **9%**.
- ❖ **Data Science Intern, Innodatatics.ai** – Pune, India Dec 2020 – Sept 2021
 - Built and deployed supervised classification models (Python, Scikit-learn) for manufacturing analytics, automating operator performance scoring; handled feature engineering, model evaluation and hyperparameter optimization, and shipped Streamlit inference apps for business users.

AI System Design & Architecture

- **RAG Systems:** chunking strategy (recursive, semantic, parent-document), hybrid retrieval, reranking, query rewriting and decomposition, metadata filtering, grounding and citation enforcement, context-window and cost budgeting.
- **Agentic Systems:** tool-use agents, ReAct and plan-and-execute patterns, supervisor/router orchestration, state and memory design, retry and failure handling, loop and cost guardrails.
- **LLM Serving & Cost:** deployment topologies on AKS and Azure ML endpoints, streaming, prompt and semantic caching, model routing, token-cost and latency optimization, PTU vs. pay-as-you-go capacity planning.
- **Evaluation & Reliability:** golden datasets, faithfulness / answer-relevancy / context-precision metrics, LLM-as-judge, regression suites in CI, hallucination detection, drift and decay monitoring.
- **Data Architecture:** medallion (bronze/silver/gold) design, Kimball star-schema modelling, SCD Type 1/2, incremental and idempotent loads, Delta Lake ACID and time travel.
- **Secure Enterprise AI:** Managed Identity, Key Vault, scoped RBAC, private networking, responsible-AI guardrails and content safety, auditability for regulated environments.

Skills

Languages & Query: Python, PySpark, SQL (window functions, CTEs, query tuning), DAX, Bash

Generative & Agentic AI: Azure OpenAI (GPT-4/4o, o-series), Llama, Mistral, BERT; Retrieval-Augmented Generation, prompt engineering (few-shot, chain-of-thought), structured output & function calling, fine-tuning (LoRA, QLoRA, PEFT); LangChain, LangGraph, LlamaIndex, CrewAI, Hugging Face Transformers; Azure AI Search, ChromaDB (HNSW indexing, hybrid search)

LLMOps & MLOps: CI/CD for ML (Azure DevOps, GitHub Actions), MLflow tracking & model registry, model serving (Docker, Kubernetes, AKS, Azure ML managed endpoints), blue-green & canary rollout, drift detection (PSI, KS), prompt versioning, evaluation-in-CI, observability & tracing, Azure AI Content Safety

Machine Learning: Regression, classification, ensembles (Random Forest, XGBoost, LightGBM), time-series forecasting (ARIMA, Prophet, feature-based ML, walk-forward validation), NLP (text classification, NER, topic modelling), transformers; imbalanced learning, probability calibration, leakage prevention, ROC-AUC / PR-AUC / KS / Gini, SHAP explainability, A/B testing

Microsoft Azure: Azure Machine Learning, Azure AI Foundry, Azure OpenAI Service, Azure AI Search, Azure AI Services, Azure Databricks, Azure Data Factory, Azure Synapse Analytics, Azure Data Lake Storage, Azure Functions, Azure Kubernetes Service (AKS), Azure Key Vault, Azure DevOps

Microsoft Fabric & Data Engineering: Fabric (OneLake, Lakehouse, Warehouse, Data Pipelines, Dataflow Gen2, Notebooks, Direct Lake semantic models, Real-Time Intelligence), Power BI, Apache Spark, Delta Lake, Unity Catalog, ETL/ELT design, data modelling, data governance, REST API integration

Tools & Platforms: Git, Docker, Kubernetes, Linux/Unix, Streamlit, FastAPI, pytest, JIRA, Confluence

Domain: Healthcare & Public Health (WHO), Financial Services (credit risk, ECL), Retail & CPG (demand forecasting, inventory optimization)

Selected Projects

- **Enterprise RAG Knowledge Assistant** Hybrid retrieval over Azure AI Search with reranking, citation-grounded generation, and a RAGAS-style eval harness (faithfulness, context precision/recall) wired into CI.
- **Multi-Agent Task Automation System** LangGraph state-graph orchestration with tool-use agents, checkpointed state, human-in-the-loop interrupts and budget guards; benchmarked against a single-agent baseline.
- **End-to-End MLOps Pipeline with Drift Monitoring** Automated retraining and gated promotion on Azure ML with MLflow registry, containerised serving on AKS, and PSI/KS drift alerting with rollback.
- **LLM Fine-Tuning & Evaluation Framework** LoRA/QLoRA fine-tuning with 4-bit quantization and a held-out eval suite comparing fine-tuned vs. RAG-only vs. base model.

Education & Certifications

Dr. D. Y. Patil Vidyapeeth, Pune – MBA, Artificial Intelligence & Machine Learning

June 2022 – Sept 2024

Savitribai Phule Pune University – Bachelor of Engineering

Aug 2017 – Dec 2020

Certifications: Microsoft Certified: Fabric Data Engineer Associate (DP-700) | Fabric Analytics Engineer Associate (DP-600)