

Jeril Kuriakose

PRINCIPAL DATA SCIENTIST

Riyadh, Saudi Arabia

📞 (+966) 509514280 | ✉️ jerilkuriakose10@gmail.com

(Saudi Arabia Premium Resident)



Profile / Summary Statement

Principal Data Scientist with deep expertise in **LLMs**, NLP, and **agentic AI systems**. Leads end-to-end development of high-impact platforms from data engineering and training to **MLOps**, **inference optimization**, and **secure deployment at scale**. Proven record of measurable business outcomes (productivity savings, accuracy gains, and risk reduction) and hands-on leadership of cross-functional data teams.

Recent focus: **Arabic LLMs (ALLaM)**, **agent orchestration**, curriculum **SFT**, **DPO**, **RL-SFT** hybrid training, and large-scale inference on **Kubernetes** with **vLLM/TGI/Triton**, Ray, and Azure.

Key Skills & Areas of Expertise

ML/AI	LLMs, NLP, Agentic systems, RAG, Generative AI, Time-series, Graph ML
Frameworks	PyTorch, Transformers, TRL, vLLM, TGI, Triton, Ray, LangChain, LangGraph, MLflow, DeepSpeed, Megatron-LM, NVIDIA NeMo (NeMo-RL), FastChat, Data-Juicer, AutoGen, Polars, Plotly
MLOps/Infra	Docker, Kubernetes, Azure, DVC, FastAPI, CI/CD, Dify, Monitoring (LangFuse)
Data/Storage	PostgreSQL, MongoDB, ClickHouse, DuckDB, Qdrant (Vector DB), Elasticsearch
Programming	Python

Professional Experience

Saudi Data & AI Authority (SDAIA)

Riyadh, Saudi Arabia

Principal Data Scientist (Gen AI)

Jan. 2024 – Present

- Lead data processing and inference optimization for **ALLaM (Arabic LLM)**; managed training/inference stacks with PyTorch, Transformers, TRL, **vLLM/TGI/Triton**, Ray, Kubernetes, Azure.
- Processed **50TB** RedPajama-Data-v2 via Data-Juicer; implemented **LLaMA-2** pretraining strategies (from-scratch and continual).
- Developed **generative data-cleaning** models to improve pretraining corpus quality and downstream performance.
- Fine-tuned ALLaM for **function/tool calling** and **multi-agent orchestration** in secure government environments.
- Architected **self-healing** agentic data pipelines with autonomous error recovery and optimization.
- Built custom **MCP servers** to integrate in-house data platforms and standardize tool connectivity.
- Implemented agentic **RAG** and multi-agent planning/orchestration using LangChain, **LangGraph**, and Dify.
- Deployed/optimized **ALLaM 7B/13B/70B** on **Kubernetes**; identified optimal inference parameters across use cases.
- Designed **curriculum SFT** pipeline for agentic capabilities; improved model convergence by **35%**.
- Developed **interleaved RL** with periodic **SFT** injections to stabilize agent policy optimization.
- Created iterative **DPO** framework with uncertainty-aware preference collection and dynamic reward modeling for reasoning.
- Introduced self-optimized fine-tuning (**SOFT**) via self-distillation and adaptive curriculum scheduling.
- Applied gradient separation to prevent interference between RL and SFT objectives in multi-task settings.
- Evaluated Grok, DBRX, Command R+ using **LM-Harness** for English and Arabic; informed model selection and deployment.
- Integrated **Dify** for function calling/agent capabilities; leveraged **LangGraph** for tool-calling description optimization.
- Core contributor to **ALLaM PC** (LLM-on-laptop) efforts enabling constrained-device usage.
- Deployed high-throughput, low-latency **Kimi-K2** and **GLM** models (v4.5–4.7) for production inference.
- Built a **user analytics platform** for LLM chat telemetry using **ClickHouse** (ingest/OLAP), **DuckDB** (ad-hoc), **Polars**, and **Plotly**; delivered product insights and quality metrics.

Mizuho Bank

Senior Data Scientist

Singapore

Jul. 2019 – Dec. 2023

- **Phoenix**: generic NLP IE platform using BERT + hybrid CNN + Bloom embeddings; **90%** accuracy; saved **15,500 man-hours/year**.
- **Phoenix**: LightGBM post-prediction for missing keywords; LSTM attention to correct OCR typos; BART-based spell checking.
- **PIGEON**: document classification with RoBERTa and summarization (GPT-2/GPT Neo); achieved **97%** accuracy; full API productionization.
- **SWAN**: AML name check reduced false positives by **20%**; vector DB for embeddings; graph neural networks for anomaly detection.
- **HAWK**: hybrid encoder-decoder LSTM for daily/weekly/monthly forecasts; saved **3,000 man-hours/year**; external data ingestion + ELK.
- **Oxygen**: hybrid Seq2Seq + BERT extraction for loan agreements; saved **6,500 man-hours/year**.
- **Sophia**: hybrid BERT + InferSent internal chatbot; built core response-prediction model.
- LLM PoCs: Falcon-7B, LLaMA-2-7B fine-tuned via **QLoRA/PEFT/bitsandbytes** on dual T4 GPUs; explored multi-modal use cases.
- Introduced **MLflow** for org-wide model tracking; implemented ML monitoring and performance validation in production.
- Adopted **Ray**/Ray AIR for distributed training and multi-GPU serving; standardized **FastAPI** microservices for deployment.
- Built internal PyPI and **Docker Registry** to accelerate packaging and image distribution across the bank network.
- Created reusable pipelines for RPA, anomaly detection, and customer segmentation to improve team productivity.
- Designed object-oriented codebases, CI/CD practices, and model governance workflows; led architecture for AI projects.
- Established **active learning** loops to auto-annotate large datasets and continuously improve model quality.
- Mentored junior data scientists; peer-reviewed AI work; tracked research trends to guide adoption.

Baker Hughes (GE Company)

Senior Data Scientist

Kochi, India

Jan. 2019 – Jun. 2019

- Predicted **stuck-pipe** events with feature-engineered ensemble models; identified key drivers and timing of incidents.
- Analyzed non-productive time (**NPT**) and drill-bit wear; delivered actionable insights to reduce rig downtime.
- Drove awareness of applied data science among SMEs through targeted training.

Innovation Incubator

Senior AI Engineer

Thiruvananthapuram, India

Feb. 2018 – Jan. 2019

- **OCR** text extraction from images using **Pytesseract**; axis detection and Excel export for downstream keyword extraction.
- Generic keyword extraction across PDFs/TXT/Excel/HTML via ensemble ML and feature engineering; production APIs.
- Document-type identification using ensemble clustering; automated routing of RCM documents.
- Customer segmentation and claim-settlement prediction; improved healthcare RCM by 18%.
- Owned end-to-end ML pipelines and backend services; collaborated with customers to translate business to ML problems.
- Mentored client teams on data and analytics best practices.

Raw Data Technologies

Senior ML/AI Developer

Kochi, India

Apr. 2017 – Feb. 2018

- EMI defaulter prediction with ensemble models; extensive feature engineering; deployed APIs and data pipelines.
- Performance analysis of internal vs external collection agents using **Kaplan-Meier** estimator and cumulative incidence.
- Logistics: demand forecasting; promotional price optimization; markdown optimization for long-lifecycle container rentals.
- Research project: **evolutionary algorithms** + **Hyperopt** to optimize automotive design structures.

St. John College of Engineering and Technology

Assistant Professor & Data Analyst (summary)

Palghar, India

Jun. 2015 – Mar. 2017

- Taught ML/Data Mining; developed analytics and ERP modules; produced institutional reports and student performance analytics.

Manipal University Jaipur

PhD Scholar, Developer, Data Analyst (summary)

Jaipur, India

Jul. 2013 – Jun. 2015

- Maintained university ERP; conducted predictive analytics on student outcomes; built data pipelines and departmental reporting.

Education

Manipal University Jaipur, School of Computing and IT

Ph.D. in Computer Engineering

Jaipur, India

Jul. 2013 – Dec. 2019

- CGPA: 9.62
- Research focus: Secure localization and malicious node detection in wireless/ad-hoc networks.

ISIM, University of Mysore

M.Tech. in Information Technology

Mysuru, India

Jul. 2012

- Graduated with 62%
- Thesis: Localization in Wireless Networks in the presence of Cheating Beacon Nodes.

Jeppiaar Engineering College (Anna University)

B.Tech. in Information Technology

Chennai, India

May 2010

- Graduated with 68%.

Selected Publications

A review of deep learning-based approaches for detection and diagnosis of diverse classes of drugs

*Archives of Computational
Methods in Engineering
(Springer)*

Kuriakose J., et al.

2023

EMBN-MANET: Eliminating Malicious Beacon Nodes in UWB-based Mobile Ad-Hoc Networks

Ad Hoc Networks (Elsevier)

Kuriakose J., Joshi S., Bairwa A. K.

2022

Secure Multipoint Relay Node Selection in Mobile Ad Hoc Networks

*Communications in Computer
and Information Science
(Springer)*

Kuriakose J., Amruth V., Raju R. V.

2015

A Review on Mobile Sensor Localization

*Security in Computing and
Communications (Springer)*

Kuriakose J., et al.

2014

A Review on Localization in Wireless Sensor Networks

*Advances in Signal Processing
and Intelligent Recognition
Systems (Springer)*

Kuriakose J., Joshi S., Raju R. V., Kilaru A.

2014

Additional publications: 30+ (full list available upon request)

Awards & Honors

Jul. 2022 **Project of the Year**, Mizuho Bank

Singapore

Oct. 2019,

May 2021, **Employee of the Month**, Mizuho Bank

Singapore

Jul. 2021

Feb. 2020 **Honourable Award – AML Project**, Mizuho Bank

Singapore

Dec. 2020 **Honourable Mention – Phoenix Project**, Mizuho Bank

Singapore

Languages

Languages English, Hindi, Tamil, Malayalam

References

References available upon request.