Aravindkumar K S

Cork, Ireland | +353 89 403 4128 | aravindkumar111333@gmail.com
linkedin.com/in/aravindkumar-ks/ | github.com/AkorNero | Stamp 1G – eligible to work in Ireland, no sponsorship required

Summary

AI/ML engineer who has built production RAG and LLM systems end to end in industry — from document ingestion, retrieval and evaluation through inference APIs to the AWS infrastructure that runs them — and went on to formalize that work with an MSc in Mathematical Modelling and Machine Learning (First Class Honours). Proficient in Python, FastAPI and LangChain, with hands-on AWS experience across ECS, Lambda, Step Functions and Terraform-based infrastructure-as-code. Takes an evaluation-driven approach, using RAGAS and Langfuse to catch regressions and ship reliable, low-risk releases, and applies solid security and compliance practices (least-privilege IAM, PII redaction, GDPR). Comfortable owning ambiguous problems from requirements through to deployed, monitored services and translating technical decisions for non-technical stakeholders. An AWS Certified Machine Learning Engineer, grounded in a mathematical and research foundation that keeps practical engineering rooted in fundamentals.

Experience

AI/ML Engineering **Oct 2025 – Present**
Independent Projects & Professional Development **Cork, Ireland**

- Designed and built production-grade AI systems to deepen applied LLM and MLOps expertise — an agentic RAG assistant, a self-hosted vLLM inference service, and a CI-integrated LLM evaluation harness (detailed under Projects).
- Earned the AWS Certified Machine Learning Engineer – Associate certification and completed LangChain and Anthropic Model Context Protocol (MCP) courses, formalizing cloud ML and agent-tooling skills.
- Mentored first-time programmers as a Django Girls volunteer and stayed engaged with the AI engineering community while pursuing full-time AI Engineer / Forward Deployed Engineer / Software Engineer roles

Senior Consultant – Package Implementation, Insurance R&D **Oct 2022 – Aug 2024**
LTIMindtree (now LTM) **Bengaluru, India**

- Led the proof-of-concept and co-built a production insurance RAG assistant (LangChain, FastAPI), cutting bulk policy-document processing time ~30% including manual extraction-quality review, at a 4–5s p95 time-to-first-token.
- Owned an event-driven extraction pipeline (Textract, S3, SQS with dead-letter queues, Lambda consumers, Step Functions) parsing 20+ vendor formats into a Neo4j knowledge-graph insurance ontology at 95%+ accuracy, auto-quarantining malformed inputs.
- Built ingestion, chunking and embedding pipelines feeding Chroma-based hybrid retrieval (dense vectors + BM25 via Reciprocal Rank Fusion), sustaining bulk re-index throughput through SQS-buffered Lambda workers, with versions tracked in MLflow.
- Co-built Studio, the team's central model portal (React, FastAPI) providing model access, management and governance — rate limiting, PII redaction, content moderation, domain gating and application-based access control.
- Independently built a VS Code extension on Studio bringing LLM chat, code generation, review and editing, model switching and history into the IDE for the engineering team.
- Contributed to the LLM inference gateway behind Studio (FastAPI edge, internal gRPC/protobuf), adding a Redis semantic cache that cut redundant LLM calls ~5% and helping enforce GDPR-compliant PII handling.
- Contributed to evaluation baselines using a RAGAS harness and Langfuse tracing, wiring context-relevance and faithfulness checks into CI as release gates to catch retrieval regressions.
- Contributed to migrating infrastructure-as-code from CloudFormation to Terraform, building reusable ECS/ECR/Lambda/S3 modules with least-privilege IAM and encryption, and cutting AWS spend ~10% through ECS right-sizing and autoscaling.
- Contributed to containerized blue-green ECS releases via GitHub Actions and CodeBuild, with SonarQube and Trivy security gates enabling low-risk deployments across Dev, Test and Prod.

Graduate Engineer Trainee **July 2022 – Oct 2022**
LTIMindtree (now LTM) **Bengaluru, India**

- Built Angular tooling for configuring data-pipeline scan jobs and reviewing classification results and privacy insights, giving analysts a self-serve interface to manage scans without engineering support.
- Designed PostgreSQL data models with Spring Boot and Hibernate ORM for data sources, job-run history and classified elements, supporting reliable tracking of scan lineage across runs.
- Wrote SQL queries and aggregations powering data-discovery and coverage-monitoring dashboards, giving stakeholders visibility into classification coverage across data sources.

- Delivered maintainable features in sprint increments through Agile cycles and peer reviews, improving code quality and reducing review turnaround.

Projects

Agentic RAG & Conversational Knowledge Assistant

Python, Gradio, FastAPI, LangChain, OpenAI, Pinecone, Redis, Pydantic, RAGAS

- Built an agentic RAG assistant where a LangChain tool-calling agent routes queries across vector search, SQL lookup and web search, returning Pydantic-validated structured outputs over multi-turn Redis-backed conversation state.
- Tuned chunking (semantic vs fixed-size, overlap) and hybrid search using OpenAI embeddings and Pinecone, diagnosing retrieval failures like missed context through RAGAS context-recall analysis.
- Added an LLM-based response classification and filtering guardrail layer, improving answer grounding and quality by catching ungrounded or irrelevant responses before return.

Serverless Observability on AWS

Terraform, AWS (Lambda, API Gateway, SQS, DynamoDB, CloudWatch, SNS, KMS), Python

- Built a modular Terraform stack for a serverless order service with sync (API Gateway → Lambda → DynamoDB) and async (SQS → worker Lambda + DLQ) paths on an S3-locked remote backend.
- Implemented structured JSON logging, metric filters and EMF custom metrics, with saved Logs Insights queries for latency percentiles, cold starts and request tracing.
- Applied least-privilege IAM, PII masking via CloudWatch Logs data protection, and KMS-encrypted SNS alerting with scoped key and topic policies.
- Added fault injection (error rate, latency, poison messages) to reproduce failures and validate metrics and alerting end to end.

Wendland's CSRBF in Kolmogorov–Arnold Networks (KAN)

Research Project – University College Cork, under Dr. Andreas Amann

- Investigated efficient Kolmogorov–Arnold Network (KAN) alternatives by replacing B-spline and Gaussian edge activations with compactly supported Wendland radial basis functions, implemented in PyTorch with trainable centers and support radii.
- Matched FC-KAN accuracy (97.6% vs 97.7% on MNIST) with 16% fewer parameters and 21–23% faster training, and improved Fashion-MNIST accuracy from 86.7% to 88.7% across six baselines.
- Reduced activation density to ~25% (vs 88% for Fast-KAN) using fixed radii, and applied inverse-L2 regularization to stabilize trainable radii, quantifying the sparsity–accuracy trade-off.

Predictive Modeling for Customer Churn

Python, pandas, scikit-learn, XGBoost, Optuna, SHAP, MLflow

- Built an end-to-end churn-prediction pipeline on 50k+ customer records, engineering behavioral and temporal features and handling class imbalance with SMOTE and stratified cross-validation
- Trained and compared logistic regression, random forest and XGBoost models, tuning hyperparameters with Optuna to reach 0.89 ROC-AUC, with experiments tracked in MLflow.
- Explained predictions with SHAP to surface top churn drivers, translating findings into actionable retention recommendations communicated to non-technical stakeholders.

Optimized LLM Inference Serving

Python, vLLM, FastAPI, PyTorch, Hugging Face Transformers, Docker, Prometheus

- Deployed a self-hosted LLM inference service with vLLM behind a FastAPI gateway, using continuous batching and PagedAttention to raise throughput ~3x over a naive Hugging Face pipeline.
- Applied INT8/AWQ quantization and KV-cache tuning to cut GPU memory ~50% and p95 latency ~40%, enabling larger models to fit on a single GPU.
- Benchmarked throughput, latency percentiles and tokens/sec under concurrent load, exposing Prometheus metrics to drive autoscaling and catch performance regressions.

LLM Evaluation Harness

Python, pytest, RAGAS, DeepEval, Langfuse, OpenAI/Anthropic APIs, GitHub Actions

- Built a reusable LLM evaluation harness that scores models and prompts on curated datasets across accuracy, faithfulness, relevance and latency, with pluggable metrics and provider-agnostic adapters.
- Implemented LLM-as-judge and reference-based scoring with RAGAS and DeepEval, calibrating judge prompts against human labels to improve metric–human agreement.
- Wired the harness into CI as regression gates via GitHub Actions, blocking merges on score drops and logging per-run traces to Langfuse for diagnosis.

Skills

AI & LLM Engineering: RAG & Agentic RAG, LangChain, tool/function calling, prompt engineering & versioning, LLM guardrails (content moderation, PII redaction, domain gating, rate limiting), structured outputs (Pydantic), LLM inference serving & optimization (vLLM, continuous batching, PagedAttention, KV-cache, INT8/AWQ quantization), semantic caching, OpenAI & Anthropic APIs, Model Context Protocol (MCP).

Retrieval & Data: Vector search (Pinecone, Chroma), graph retrieval (Neo4j), hybrid search, chunking strategies, embeddings, knowledge graphs & ontologies, document extraction (Amazon Textract), Redis, PostgreSQL, DynamoDB, data modelling.

Coding Agents & Agentic Development: Claude Code, OpenClaw, Hermes, Herdr, agent skills & plugins, hooks, subagents, MCP servers & tool integration, Docker-sandboxed agent execution, agentic coding workflows, greenfield & brownfield development

Cloud: AWS (ECS, ECR, Lambda, API Gateway, S3, SQS & DLQ, DynamoDB, Step Functions, CloudWatch, SNS, KMS, EFS).

MLOps & Security: Terraform (IaC), CloudFormation, Docker, CI/CD (GitHub Actions, CodeBuild), blue-green deployments, MLflow, RAGAS, Langfuse, DeepEval, evaluation harnesses & CI regression gates, SonarQube, Trivy, least-privilege IAM, KMS encryption, PII masking, CIS Benchmarks, GDPR compliance, observability (Prometheus, Grafana).

Architecture & System Design: Event-Driven, Microservices & Serverless Architecture, REST & gRPC (Protocol Buffers), sync/async messaging, dead-letter queues, caching strategies, SOLID Principles, Least Privilege & RBAC, LLMOps Lifecycle, Spec-Driven Development (SDD), Test-Driven Development (TDD), Distributed Tracing, Fault Injection Testing, Green and brown-field projects.

Machine Learning: Deep learning (ANN, RNN, VAE, Transformers, RL), Kolmogorov–Arnold Networks, Radial Basis Functions, Regularization, Statistical Modeling, EDA, feature engineering, hyperparameter tuning (Optuna), class imbalance handling (SMOTE), model interpretability (SHAP), model evaluation & benchmarking.

Languages & Frameworks: Python, FastAPI, React, SQL, PyTorch, scikit-learn, XGBoost, pandas, Java, JavaScript/TypeScript, Gradio, Flask, Spring Boot, Hibernate, Angular.

Education

MSc in Mathematical Modelling and Machine Learning — **Sept 2024 – Sept 2025**
University College Cork, — **Cork, Ireland**

- First Class Honours (1.1)
- Coursework: Deep Learning (ANN, RNN, VAE, Transformers, RL), Statistical Methods (Data Analysis, EDA, Regression, Classification, Unsupervised Models).

BE in Mechanical Engineering — **July 2018 – June 2022**
Kumaraguru College of Technology, — **Coimbatore, India**

- First Class with Distinction (8.8/10.0 CGPA)
- Publication: “Design Optimization of Grids in Grid Fins for Spacecrafts”, International Journal of Vehicle Structures and Systems (IJVSS), 2022.

Certifications

AWS Certified Machine Learning Engineer – Associate – Amazon Web Services	Sept 2026
Foundation - Introduction to LangChain - Python – LangChain	May 2026
Model Context Protocol: Advanced Topics – Anthropic	April 2026
Introduction to Model Context Protocol – Anthropic	March 2026

Awards & Volunteering

Mentor (Volunteer) – Django Girls, Cork — **Sept 2026**

- Coached a group of 3 first-time programmers through building and deploying a full Django blog application to PythonAnywhere in a single-day workshop.

Shooting Star Award – LTIMindtree, Bengaluru — **April – June 2024**

- Recognised as a top team contributor for consistently delivering high-impact features on the Studio platform during Q1.

Rising Star Award – LTIMindtree, Bengaluru — **Sept 2023**

- Young Achiever award for exemplary contribution in my first year at the company.