

Ritesh Yadav

✉ daydreamingguy941@gmail.com | ☎ 6386461149 | Kanpur

AI/ML Engineer | 5.5+ Years of Experience

🐙 github.com/DARK-art108 | 🔗 [linkedin.com/in/ritesh-yadav2011/](https://www.linkedin.com/in/ritesh-yadav2011/) | 📄 <https://www.kaggle.com/ritesh2000> |
👤 <https://codeforces.com/profile/q1w2e3r4t5>

Skills

Languages: C/C++, Advance Python

Database: MongoDB , Mysql, PostgreSQL

Machine Learning: Scikit-learn , Dask , Statistics, PySpark

MLOps: Docker, Mlflow, Kubeflow, Multi Model Server, Kubernetes, Github Actions, Jenkins, AWS, Azure, Minio, Sonar-Qube, Prisma Cloud, Keda, Helm, Prometheus, Apache Airflow, Spark, Kafka, Snowflake, Hadoop, Hive, SGLang, vLLM, KV-cache Optimization, Model Quantization

Rest-API: FastApi , Flask

Deep Learning: Pytorch, Tensorflow

Natural Language Processing: Classification, Summarization, Text Generation, Name Entity Recognition, Transformers, Retrieval Augmented Generation, Retrieval, GraphRAG, Langchain, Llamaindex, FastMCP, Ollama, Groq, PydanticAI, MCP, CrewAI, Google ADK, Google Vertexai

Work Experience

66degrees, Remote

July 2025 - Present

AI/ML Engineer

Demand & Price Forecasting Platform (J.D. Power)

- Architected a multi-horizon time-series forecasting system using **TFT (Temporal Fusion Transformer)** and **TiDE** as champion-challenger models across **141 model configurations** spanning daily, weekly, and monthly horizons.
- Engineered segment-cluster model grouping across **9 clusters plus 1 all-segments model**, reducing computational overhead while preserving forecast granularity at **Segment, DMA, and MSF** levels.
- Implemented **quantile regression** for TFT and TiDE models to generate prediction confidence intervals and improve forecast reliability for business planning.
- Built an explainability meta-model and waterfall attribution pipeline to quantify period-over-period feature importance shifts, identifying statistically significant drivers such as **relative incentives, dealer margin, and events**.
- Designed a scenario simulation engine supporting multi-covariate future adjustments, resolving normalization and future-period adjustment defects that caused counter-intuitive forecast outputs.
- Refactored training orchestration to submit parallel jobs to **Vertex AI**; integrated **MLflow**, Parquet-based feature importance storage, hierarchical reconciliation, and **GKE** dashboard deployment with **Okta** authentication.

ARC Implementation Agents (RealPage)

- Designed multimodal document extraction for **PDFs, scanned images, CSV, and XLSX** using **Gemini on Vertex AI** and deterministic tabular pipelines, with Pydantic-validated outputs persisted to **GCS** and **Firestore**.
- Built gold-standard evaluation harnesses, LLM-as-judge quality checks, extraction and segmentation eval runners, and per-call token/cost telemetry for property-management conversion workflows.
- Implemented a provider-agnostic **LLMOps** layer supporting **Gemini, LiteLLM, and Google ADK**, structured JSON outputs, centralized prompt specs, judge pipelines, and **OpenTelemetry/LangSmith** tracing.
- Led hub-and-spoke agentic orchestration using a **LangGraph** supervisor and FastAPI, coordinating L1 workers through canonical work-order/result contracts with retries, idempotency, checkpointing, and human-in-the-loop gates.
- Built event-driven integrations connecting **Salesforce CDC, DataHub, SendGrid, GCS, Firestore, and Outlook** to multi-agent workflows, including leader-elected CDC and idempotent CSR synchronization.
- Shipped cloud-native services on **GKE** across dev, UAT, and production using **Helm**, GitHub Actions, Workload Identity, HPA, Gateway API HTTPRoutes, and production observability.

Enterprise AI & Workflow Automation (Boston Dynamics)

- Designed and deployed enterprise AI agent workflows and system prompts on **Gemini Enterprise** to automate complex operational and technical tasks using **Google Knowledge Graph Search Integration** and **Vertex AI Search**.
- Configured secure, multi-platform federated connectors and **OAuth-authenticated REST API integrations** across **Google Workspace (Drive, Gmail, Calendar, Docs, and Slides)**, **Jira Cloud/Jira Service Management**, **Confluence Cloud**, and **Salesforce CRM**.

- Implemented multi-step AI agents to retrieve technical documentation, synthesize insights, generate Google Docs, and create or update Jira backlog items and sprint-review tasks across connected enterprise systems.
- Built automated reporting workflows that extracted resolution details from closed Jira issues into Google Slides templates and analyzed meeting transcripts to generate agendas for upcoming Google Calendar events.
- Developed Salesforce AI workflows to validate Sales Orders against Purchase Orders and Quotes, reducing fulfillment and invoicing errors; implemented rule-based customer-support case triage across opportunity and account data.
- Defined Google Workspace Admin security policies, OAuth controls, data-store boundaries, IAM access rules, folder targeting, and chunking parameters; performed iterative prompt tuning, Build Verification Testing, UAT, API load/stability checks, and retrieval-accuracy validation.

IntellectAI, Remote

July 2023 - June 2025

MLE & MLOps

- Designed and scaled a complex **GenAI platform** for leading financial firms to process, extract, and classify millions of diverse financial documents, with **real-time data streams** and scalable knowledge bases, achieving **latency under 200ms, 97.9% uptime**, and seamless horizontal scaling.
- Designed and integrated the **GenAI Conversation, RAG and Multi-Agent workflows** for the platform with realtime tracing using **Langfuse**.
- Designed and implemented an automated evaluation module to assess **LLM and LMM-generated responses** across **Correctness, Faithfulness, and Relevancy metrics**, leveraging **GPT-4 and GPT-4o-based evaluators (LLM-as-a-Judge and Multi-Modal LLMs)** for comprehensive quality assessment in RAG workflows.
- Developed and integrated an MVP on **Reasoning and Action Agents, Chain-of-Thought (CoT) and Tree-of-Thought (ToT)** using Langchain.
- Enabled Agent remote code execution, custom callbacks, file upload, Tool Selection.
- Worked on a PoC of **audio KB pipeline** using **OpenAI-Whisper, PyAudio, and Pydub**.
- Hosted real-time, asynchronous, and serverless endpoints for **Hugging Face models on SageMaker**.
- Implemented queue-based, **event-driven scaling** for the embedding generation service to reduce costs.
- Designed and implemented a **Model monitoring pipeline** for Deep Learning workflows using **Evidently**.
- Developed a wrapper to attach **AWS Guardrails** on top of the GenAI Platform for **toxicity, PII information detection**.
- Implemented **dynamic affinity and toleration settings** for CPU and GPU workloads on Kubernetes for **dynamic routing**.
- Reduced the build and runtime of the platform's ML backend tools **Kubeflow pipeline, MLflow, Multi-model server, Jenkins**.
- Implemented **multi-stage Docker builds** to reduce build time and image size.
- Maintained a scalable **MLOps observability** stack using **Prometheus** for time-series metrics collection and **Grafana** to monitor infrastructure health, **latency (p95 < 200ms)**, **token usage (5M tokens/day for 500+ page documents with complex tables)**, and LLM costs. Achieved a **25% reduction in inference costs**, improved processing throughput by **40%**, and ensured **98.9% uptime across multi-region deployments** handling **100K+ API** requests daily.
- Implemented multi-tier monitoring with node-level observability using **Node Exporter and cAdvisor**, and custom metrics collection using Prometheus client libraries.
- Optimized **time-series data storage and query performance by deploying Thanos for long-term storage of Prometheus metrics across highly distributed environments**, ensuring data deduplication, global querying, and high availability.
- Deployed and managed **Loki for centralized, scalable log aggregation**, enabling efficient log storage and querying alongside metrics in Grafana dashboards.
- Deployed open-source LLMs (**LLaMA 3.2, Mistral 7B**) for **serverless inference using Knative**, including cold start optimizations through **pre-warmed pods**.
- Implemented **Kubernetes Admission Controllers to enforce GPU/CPU resource quotas** and validate model deployments in a GenAI MLOps pipeline, optimizing cluster resource utilization and ensuring operational efficiency.
- Designed and deployed a **Multi-Agent GenAI Chatbot** for a leading international banking client's **FX pricing platform**, utilizing **LangGraph, LangChain, and MCP(Model Context Protocol)**. The system enabled **multi-turn conversations** with dynamic task routing across specialized agents, including retrieval, planner, and fallback, streamlining complex FX pricing queries and enhancing user engagement.
- Integrated Smart **Retrieval-Augmented Generation (RAG)** with short term and long term memory using **Redis and Cosmos DB** to improve contextual awareness and user experience in financial service dialogues, while also implementing **Cache Augmented Generation (CAG)** to enhance response speed and reduce latency by preloading knowledge into the model's context window.
- Engineered **numerical and monetary precision pipelines** by integrating deterministic function tools for **FX rate calculations**, applying **banking-grade rules (banker's rounding, decimal precision, and schema validation using JSON Schema/OpenAPI)**.
- Built a **zero-trust security model** by implementing a schema-driven safety layer that rejects hallucinated responses, ensuring that only validated FX data from the **on-prem FXPCB SQL system (migrating to Azure)** is returned to the

LLM, with strict access control and continuous validation at each interaction.

- Established a full CI/CD pipeline for chatbot deployment with production-grade monitoring using Azure Monitor, Log Analytics, and App Insights, enabling real-time health tracking and latency monitoring.
- Implemented LangGraph-native tracing with LangSmith, enabling LLMOps observability on agent performance, hallucination rates, session analytics, and conversational accuracy.
- Engineered a **production-grade CI/CD pipeline** for chatbot deployment, integrating **Azure Monitor**, **Log Analytics**, and **Application Insights** for real-time health, latency, and performance tracking. Enhanced **LLMOps observability** through **LangGraph-native tracing** with **LangSmith**, enabling detailed insights into **agent performance**, hallucination rates, session dynamics, and conversational accuracy.
- Led **secure and scalable agent deployment** with **secret-less authentication**, integrating **Azure AD**, **OAuth2**, **Entra Verified ID**, and **managed identity** aligned with the **least privilege principle**. Continuously improved agent effectiveness through **rigorous real-world user testing** and iterative fine-tuning via **Reinforcement Learning from Human Feedback (RLHF)** and feedback loop analysis.

iNeuron.ai, Remote

June 2021 - June 2023

MLOps Engineer

- **Designed and deployed a scalable machine learning system to enhance fraud detection, user recommendations, and customer experience for a blockchain-based financial platform**, simulating real-world applications. The project focused on risk management, user engagement, and regulatory compliance, leveraging advanced ML techniques, deep learning, and LLMs in a production environment.
- Reduced fraudulent transactions by 35%, saving approximately \$2M annually in potential fraud-related losses.
- Enhanced platform trust and security, boosting user onboarding rates by 20%.
- Delivered an end-to-end ML solution adopted across multiple internal teams, democratizing ML usage and improving time-to-market for new features.
- Migrated ML workflows to native AWS services, enabling automated scalability in response to workload demands and enhancing logging and troubleshooting processes.
- Built an **end-to-end server-side architecture for real-time stock market prediction** using various tools & libraries.
- Python, AWS Step Functions, AWS Batch, Lambda, S3, DynamoDB, EC2, SQS, SNS, AWS CDK, AWS Athena, Elastic Search, LightGBM, TensorFlow, Graph Neural Networks (GNN), Fraud Detection, Recommendation Systems, Risk Assessment, Blockchain Mining, Scalable Backend Systems, Data Processing, MLOps, Production Deployment, AWS, Kubernetes, Docker, CI/CD, Apache Spark, Airflow, Feature Engineering, Data Mining, Information Retrieval, Natural Language Processing (NLP), LLMs, GNN, Deep Learning, Logistic Regression, Gradient Boosting trees

Education

Jaypee University of Engineering and Technology, Guna

July 2019 - May 2023

Bachelors of Technology in Computer Science and Engineering

CGPA: 7.2/10

Relevant Coursework: Object Oriented Programming, Databases, Discrete Maths, Data Structures and Algorithms, Operating Systems, Computer Networks, Machine Learning, Data Mining, Advance Data Structures and Algorithms, Information Retrieval, Image Processing

Project Work

- **Reverse Search Engine (2022):** Architected and deployed a **microservices-based image search engine** leveraging **ResNet18** for embedding generation and **ANNOY** for **low-latency approximate nearest neighbor retrieval**. Integrated **AWS S3** as a scalable data store with controlled public ACLs for secure image access post-inference. Implemented host-controlled continuous training pipelines using **GitHub Actions**, with models versioned and managed via an S3-based registry. Automated model reloading at inference endpoints through **AWS Lambda triggers**, ensuring seamless updates and high availability of prediction services. Python, AWS S3, Github Actions, AWS Lambda, Pytorch, EC2, ECR, GitHub Actions MongoDB, Docker, PaperSpace.
- **Megatron ChatBot (2021):** Megatron is a closed-domain Chatbot trained from scratch to answer various user queries. The ChatBot is embedded on the website and accept the query in various format's like audio & multilingual text. Python, Pytorch, Hugging Face, FastApi, Docker, ECR, Terraform

Awards and Certificates

- Certified Kubernetes Administrator (**CKA**)
- Certified Kubernetes Application Developer (**CKAD**)
- Google Cloud Certified Professional Machine Learning Engineer (**PMLE**)
- Google Cloud Certified Generative AI Leader
- Google Certified TensorFlow Developer
- Awarded title of **Kaggle Notebook Master**