

Abhay Raj Singh

Associate Architect - Generative AI

Toronto || Canada || +1 (416) 258-9095 || +91 6395678720 || raj.abhay6@gmail.com

<https://www.linkedin.com/in/abhay263/> || <https://github.com/Abby263>

HIGHLIGHTS

- Experienced Data Scientist Leader with over 9 years of expertise in building impactful Machine Learning (ML) solutions.
- Proficient in various ML techniques including Generative AI, Deep Learning, Natural Language Processing, and building cutting-edge Large Language Models (LLM) applications using Python and ML Frameworks like PyTorch and TensorFlow.
- Expertise in customized high-performance AI accelerator/GPU and infrastructure design and architecture.
- Acknowledged and swiftly promoted for enhancing efficiency, accuracy, and scalability of AI/ML models deployed into production.

SKILLS

Programming Languages: Python, SQL, JavaScript, SAS, R

AI & ML Frameworks: PyTorch, TensorFlow, Keras, Hugging Face, Scikit-learn, OpenCV, Whisper, ONNX, XGBoost, LightGBM

Generative AI & Agentic Frameworks: LangChain, LangGraph, Langsmith, LlamaIndex, CrewAI, Azure Foundry SDK, LLM Fine-tuning, Prompt Engineering, RAG, Agentic Memory (Short-term & Long-term), Tool Calling, Chain-of-Thought Prompting, A2A, MCP Server

Computer Vision & Edge AI: NVIDIA DeepStream, GStreamer, RTSP, Object Detection, Object Tracking, NVIDIA TAO Toolkit, Jetson (Nano/Xavier), Edge Computing, OWLViT

Speech AI: NVIDIA RIVA (ASR & TTS), Automatic Speech Recognition (ASR), Text-to-Speech (TTS), Streaming TTS, Word Boosting, Conversational AI, Virtual Avatars, NVIDIA ACE (Avatar Customization & Streaming), WebRTC

Web & App Development: FastAPI, Flask, Node.js, React.js, Streamlit, Chainlit

Cloud & DevOps: Azure (AI Foundry, Cognitive Services, Databricks, ML Studio, Blob Storage, Data Lake, Genie, DevOps), AWS (SageMaker, Bedrock), GCP, Oracle Cloud Infrastructure (OCI), Docker, Podman, Kubernetes, Rancher, Jenkins, Grafana, CI/CD, Low-Latency Inference, Apache Kafka

Databases & Vector Stores: Pinecone, Qdrant, ChromaDB, PostgreSQL, MySQL, MongoDB, Cassandra, Redis, Neo4j, GraphDB, Azure AI Search, Azure CosmosDB, Amazon Neptune, Elasticsearch, OpenSearch

NVIDIA Ecosystem: NVIDIA NIMs, NVIDIA NeMo Guardrails, NVIDIA ACE, AgentIQ

PROFESSIONAL EXPERIENCE

Associate Architect - Machine Learning
Quantiphi Analytics Canada Limited

Oct 2022 - Present
Toronto, Canada

- **Enterprise Clinical Protocol Intelligence & Agentic Search Platform**
 - **Architected a trusted clinical protocol knowledge platform** to digitize complex study protocols, validate extracted intelligence through human review, enable agentic search, and publish governed outputs across enterprise workflows.
 - **Designed a GPU-backed document-understanding pipeline** on ECS/EC2 GPU Auto Scaling Groups to generate section-aware markdown, tables, layouts, and bounding boxes, followed by LLM schema extraction with confidence scores and source citations.
 - **Built human-in-the-loop quality control** through a dual-pane PDF and extracted-field review experience, enabling users to edit, approve, and trace every field to its source before search indexing or downstream publishing.
 - **Architected Agentic Search and Agentic RAG** over a protocol knowledge base using OpenSearch embeddings and structured features, supporting natural-language Q&A, semantic retrieval, similarity search, and filterable discovery.
 - **Developed LangGraph-based deep agents** for document understanding, schema reasoning, retrieval planning, and search orchestration across unstructured protocol content and integration-ready structured study data.
 - **Enabled Schema Studio and agent-assisted extraction** so users can define and refine schemas while adaptive agents select and evaluate extraction tools to produce high-quality, field-level outputs.
 - **Delivered secure enterprise consumption channels** through protocol APIs, webhooks, and MCP exposed via API Gateway, allowing approved protocol intelligence to power analytics, regulatory workflows, and other agentic applications beyond the portal.
 - **Designed a scalable multi-tenant AWS architecture** using EKS services, Aurora, OpenSearch tenant isolation, EventBridge/SQS, S3, Secrets Manager, a peered RunPulse VPC, and Kong/LiteLLM gateways with Amazon Bedrock for governed model access.

- **Hardened production AI operations** with LangSmith and Datadog observability, policy and safety guardrails, tenant-aware access controls, and an evaluation flywheel using human feedback, LLM-as-judge, and prompt iteration.

Skill Set: Python, LangGraph, Agentic RAG, OpenSearch, Amazon Bedrock, AWS EKS, ECS, EC2 GPU ASG, Aurora, S3, EventBridge, SQS, API Gateway, Kong, LiteLLM, MCP, Webhooks, RunPulse, HITL, LangSmith, Datadog, LLM-as-Judge, Multi-Tenant Architecture, Security Guardrails

- **Agentic Automation for Financial Reconciliation & Compliance Reporting**
 - **Designed and deployed a scalable multi-agent AI system** to automate a highly manual Financial Close & Reconciliation process for a global healthcare data services firm, **reducing manual effort by 80%** and accelerating program closure timelines by over 60%.
 - Designed a **LangGraph-based orchestration workflow** with 4 modular agents (Ingestion, OCR, Reconciliation, Reporting) to automate ingestion, validation, and closure processes across **Salesforce and Microsoft SQL pipelines**.
 - Engineered a robust reconciliation agent to apply **over 100 complex business rules** spanning attendee, venue, and speaker workflows—achieving 92%+ rule-match accuracy and automating over 85% of discrepancy flagging with high interpretability.
 - **Scaled production throughput to thousands of programs per day on Azure Container Apps Jobs** through a cron-based coordinator that allocates and prioritizes new, re-triggered, recovery, version-upgrade, and Mongo backfill workloads; published work to Azure Service Bus and used KEDA-driven workers to scale from zero to approximately 50 parallel jobs.
 - **Hardened the processing fabric for production reliability** with duplicate detection, dead-letter queues, delivery retries, per-program error isolation, long-running worker timeouts, completion checks, notifications, and pipeline versioning that supports safe rule upgrades without disrupting completed programs.
 - Developed an **OCR pipeline using multimodal LLMs** to parse financial documents with 90%+ field-level extraction accuracy, validated through rule-driven reconciliation and planner feedback loops.
 - **Shipped an enterprise operations console using Next.js, Azure AD/MSAL, and Azure Container Apps** for workload triage, exception inspection, reviewer assignments, user administration, re-triggers, and human-in-the-loop compliance review, with blue-green revisions supporting controlled production releases.
 - **Operationalized a production Agent Flywheel** that converts reviewer satisfied/not-satisfied feedback into golden datasets, LangSmith prompt versions, precision/recall/F1 evaluations, and promotion gates, continuously improving reconciliation quality under real production load.
 - **Integrated LangSmith** tracing to provide real-time agent telemetry, execution audits, and explainability for business users and QA teams—shortening issue triage time by 70%.
 - Implemented robust **CI/CD pipelines on GitLab** with **Azure Container Apps, Terraform**-based infrastructure provisioning, automated preview deployments, and secure environment variable injection.
 - Designed for **high reliability with graceful error fallback logic**, intelligent rerouting, and complete audit logs, ensuring 100% downstream reporting even under partial failures.

Skill Set: Python, LangGraph, LangChain, React, Node.js, Azure SQL, Azure Cosmos DB, Salesforce, Docker, Terraform, GitLab CI/CD, Azure Container Apps, Azure Blob Storage, LangSmith, OpenAPI, YAML, Mermaid, SQLAlchemy, Pydantic, Poetry, VS Code

- **Virtual Nurse Avatar: Real-Time Voice Interaction & Intelligent Prioritization**
 - **Developed an AI-powered Virtual Inpatient Companion (VIC)** prototype leveraging Large Language Models (LLMs) for real-time patient engagement, triage, and prioritization of patient requests in a virtual command center.
 - **Designed and deployed an advanced ASR & TTS** pipeline integrating NVIDIA RIVA ASR & TTS, with hyperparameter tuning for background noise resilience, word boosting for medical terminology, and alphabet tuning for medical phrase accuracy.
 - **Engineered a robust LLM pipeline** incorporating advanced prompt engineering techniques (Few-Shot, Chain-of-Thought, Tree-of-Thought) for intent detection, slot filling, and patient query classification (Clinical vs. Environmental).
 - **Implemented LLM Guardrails using NVIDIA NeMo** with defined input/output rails to ensure safety, accuracy, and compliance in patient interactions.
 - **Developed an AI-powered prioritization algorithm** integrating LLM-based query classification with dynamic request handling in the Virtual Command Center (VN Queue).
 - **Led the development of a digital avatar interface** using NVIDIA ACE 4.0 with streaming TTS, real-time avatar rendering, and customizable styling (background, outfit, expressions).
 - **Designed and implemented an end-to-end workflow orchestration** pipeline integrating NVIDIA NIMs with ACE for real-time, scalable AI-driven patient interaction.
 - **Developed a secure, containerized deployment pipeline** on OCI (Oracle Cloud Infrastructure) ensuring scalable, low-latency AI inference for virtual patient care.
 - **Built a role-based access-controlled (RBAC)** web interface for the VN Queue, displaying real-time patient request prioritization with session management.

- **Optimized system performance for low-latency voice** interactions ensuring seamless, real-time responses with minimal delay in virtual nurse interactions.

Skill Set: LLMs, Prompt Engineering, NVIDIA RIVA (ASR & TTS), Speech Recognition, Text-to-Speech (TTS), NVIDIA NeMo Guardrails, Virtual Avatars, NVIDIA ACE (Avatar Customization & Streaming), NVIDIA NIMs, AI Workflow Orchestration, Oracle Cloud Infrastructure (OCI), WebRTC, Real-Time Streaming, FastAPI, Flask, PyTorch, TensorRT, CUDA, Edge Computing, Model Optimization, Hyperparameter Tuning (HPT), Word Boosting, Conversational AI, Message Queues (AMQP, MQTT, RabbitMQ), Streaming Text-to-Speech, React, WebSockets, Databricks (DBX), Machine Learning Operations (MLOps), Continuous Integration/Continuous Deployment (CI/CD), Low-Latency Inference.

- **NVIDIA Deepstream Multi-Camera Multi-Object CV Pipeline for Restaurant Capacity Planning with Edge Deployment**
 - Managed real-time data streams from multiple store cameras via **RTSP**, integrating interior and exterior feeds.
 - Built a robust training pipeline for object detection models using **NVIDIA TAO** and **PyTorch**, leveraging transfer learning and fine-tuning techniques to improve model accuracy.
 - Utilized the **RTDETRv2 Pytorch** model for object detection and **NvDCF Tracker** for object tracking.
 - Optimized the RTDETRv2 model through pruning, hyperparameter tuning, and other strategies to achieve $\leq 20\%$ GPU utilization on a Tesla T4 GPU.
 - Converted the model to ONNX format and experimented with FP32, FP16, and INT8 precision to balance accuracy and latency.
 - Researched and deployed **YOLOv4**, **YOLOv4-Tiny**, and **YOLOv8** models for object detection using a DeepStream application.
 - Employed the **NVDSAnalytics** plugin to gather and analyze store metadata, providing insights into customer behavior and store capacity.
 - Orchestrated the deployment of models using **NVIDIA Triton Server** for efficient inference.
 - Transmitted messages to the cloud using **RabbitMQ** Message Broker.
 - Designed and implemented an automated data labeling pipeline on Databricks utilizing the **OWLViT** model.
 - Processed detection results to determine the hold time of fries, aiding in inventory and operational efficiency.
 - Implemented Edge Deployment strategies to ensure low latency and high reliability, utilizing **NVIDIA Jetson** devices.

Skill Set: GStreamer, NVIDIA Deepstream, PyTorch, RTSP, Object Detection, Object Tracking, Transformers, CUDA, TensorRT, Transfer Learning, Edge Computing, Jetson, NVIDIA TAO, MLOps, AMQP, MQTT, RabbitMQ, Databricks, AWS, OWLViT, Kafka

- **Multi-Agent Retrieval Augmented Generation (RAG) for Educational Insights**
 - Architected and deployed a modular, scalable multi-agent system in the educational domain to deliver precise, data-driven answers to assessment-related queries, leveraging **Anthropic's Claude 3-5-sonnet** LLM on **AWS Chat Bedrock**.
 - Designed an agent workflow involving three specialized roles—**Lead Data Analyst**, **Data Engineer**, and **Visualization Specialist**—with structured goals and responsibilities for each agent, enabling nuanced understanding, data processing, and visual response generation.
 - Implemented data retrieval functionalities using **query_database**, **get_possible_answers_to_question**, and **get_questions_of_possible_type** tools to efficiently filter and analyze dataset sections, with a PostgreSQL database backend for optimized storage and query performance.
 - Configured & Integrated the visualization agent to generate insightful charts and graphs through the **generate_chart** tool, ensuring data interpretability and visualization accuracy.
 - Designed sequential processing for task-oriented delegation using a **Crew** structure, facilitating seamless inter-agent communication, data exchange, and iterative improvement up to a set threshold. Incorporated verbose logging to enhance traceability and debugging.
 - Conducted a comparative analysis by implementing the same multi-agent structure using **LangGraph** alongside **CrewAI**, evaluating performance metrics such as accuracy, response quality, cost, and token usage.
 - Utilized **AWS Elastic Container Service** and **CodeCommit** to automate deployments, ensuring smooth and efficient CI/CD processes for the system.

Skill Set: Python, CrewAI, LLMs, PostgreSQL, Matplotlib, Seaborn, Amazon Sagemaker, DBeaver, S3, Prompt Engineering, AWS Elastic Container Service, CodeCommit, LangGraph

Associate Technical Architect
Quantiphi

Dec 2018 - Oct 2022
Bengaluru, India

- **Inventory Optimization and Demand Forecasting Initiative for Quick-Service Restaurants:**
 - Optimized inventory and production management for high-demand QSR items such as filets, tenders, nuggets, and

grilled filets by building a robust forecasting solution that captures seasonality and trends, manages the machine learning model lifecycle, and generates marketing scenarios to minimize wastage and maximize fulfillment.

- Designed and implemented demand forecasting models, including exponential smoothing, seasonal naive, moving average, and stacking regression models, to optimize inventory level predictions.
- Achieved an organizational goal of maintaining less than 20% sMAPE, resulting in projected savings of \$15 million over five years.
- Built a system that uses forecasts as recommendations on batch sizes and number of Kanbans (pans) in the system to meet customer demands for production items (filets, tenders, nuggets, strips, grilled filets, etc.)
- Generated sub-hourly level point forecasts to ensure accurate, real-time demand predictions.
- Processed and handled high-volume 15-minute interval data across ~3000 store locations and multiple distribution pins.

Skill Set: Time Series, LGBM, Transformers, Amazon Sagemaker, AirFlow

- **Real-Time Alert Identification for Ride Safety in Parks using Computer Vision:**

- Implemented machine learning models for detecting various safety concerns, including guests stepping out of rides and standing during motion.
- Established a robust MLOps pipeline for continuous machine learning model integration, training, and deployment.
- Achieved high operational efficiency with a 95% true positive capture rate and less than 10 false positives per day across all cameras.
- We utilized AWS Sagemaker for model training and deployment in a cloud environment.
- We employed YOLO v4 for accurate object detection, enabling the system to identify guests and objects within the ride environment.
- Integrated 3D Convolutional Neural Networks (CNNs) for action recognition, enabling the system to classify guest behaviors such as standing during motion.

Skill Set: PyTorch, OpenCV, MLOps, Docker, Kubernetes, AWS Sagemaker, NVIDIA, Redis, Genetic, Grafana, Object Detection (YOLO v4), Image Segmentation (FPN), Action Recognition (3D CNN), Anomaly Detection (MNAD), Support Vector Classification

Analyst

Ugam Solutions, A Merkle Company

**June 2017 – Dec 2018
Bengaluru, India**

- **E-commerce Product Description Generator and Matching Tool using NLP:**

- Generated Exact and Similar Matches for the base product across multiple competitors having an accuracy of around 95% and reducing manual intervention by 90%.
- Extracted attributes from Product Images and Descriptions to create templates for writing descriptions which increased the efficiency of a manual writer by 5x.
- Integrated the solution into a Flask web interface, allowing end users to view matched products and suggested descriptions, enhancing workflow efficiency and consistency.

Skill Set: NLP, Tesseract, R, SQL, Python, Keras, OpenCV, Flask, NER, Topic Modelling, GLOVE, Word Embeddings

- **Feature-Based Opinion Mining**

- Built a high-accuracy NLP solution to analyze 1 M+ customer reviews and call logs, achieving **92% sentiment classification accuracy**.
- Extracted and categorized feature-level sentiments using GloVe embeddings, **NER**, and **Topic Modeling**, providing actionable insights for product and marketing strategies.
- Enhanced data quality through **OCR** processing with Tesseract and **OpenCV**, streamlining analysis from diverse data sources.
- Developed a real-time sentiment dashboard in **Flask**, empowering business teams to track customer sentiment shifts and improve response time, boosting satisfaction by **15%**.

Skill Set: NLP, OCR, Web Framework, Deep Learning, NLTK, EDA, NER, Topic Modelling, SQL

Data Science Intern

HP Enterprise

Image Classifier Model for MNIST Dataset

**Mar 2017 – Apr 2017
Chennai, India**

EDUCATION

Bachelor of Technology

2013 - 2017

- Vellore Institute of Technology | **Maj:** EEE | **CGPA:** 8.49/10

HSC

2013

- CBSE | **Percentage:** 87.2 %

SSC

2011

- CBSE | **CGPA:** 9.2/10