

Jestin Roy

Melbourne, VIC • +61 468364714 • jestinroy3@gmail.com • linkedin.com/in/jestinroy3 • github.com/mr-jestin-roy • jestinroy.space

SUMMARY

AI Data Scientist with **3+ years** of commercial experience, currently **working full-time as Data Scientist at Monash Business School** and holding a **Master's in Data Science** from Monash University, specializing in **LLM evaluation harnesses, AI observability pipelines, and Responsible AI guardrails** across **7+ universities & tech firms** globally. Co-founded a deeptech AI health venture (**Monash Health**) and won major awards at Monash DeepTech Entrepreneurship 2026. Focused on building high-throughput, audit-logged, and regulatory-compliant AI applications within enterprise environments. **Unlimited Work Rights in Australia.**

PROFESSIONAL EXPERIENCES

Monash University

Melbourne, VIC

Data Scientist

Jul 2025 – Present

- Architected high-performance **LLM evaluation pipelines** using **LLM-as-Judge (RAGAS)** and instrumented granular traces with **Langfuse** to continuously monitor latency, state management, and layout bottlenecks against the DocVQA benchmark (97% character accuracy).
- Co-led an **AI-assisted academic hiring and promotion platform** under Prof. Michael Ward, engineering **NLP and multi-source extraction pipelines** to synthesize candidate profiles from publications & CVs into evidence-based scores.
- Implemented **Responsible AI governance controls—fairness evaluation, bias detection, and discrimination metrics** to mitigate risk in algorithmic candidate scoring, designing output as a **human-in-the-loop decision aid** informing the first stage of faculty hiring negotiations.
- Designed and benchmarked **LangGraph multi-agent ReAct frameworks**, improving task-chain accuracy by **+18%** through tool-use optimization (MCP servers) and adaptive routing; aligning technical trade-offs with stakeholder ROI.
- Optimized 7B-parameter Vision Language Models for distributed inference on **NVIDIA A100** clusters using **10x GPU parallelization** (FSDP), resulting in maximized tokens/sec throughput, minimized cost-per-request, and a **97%** reduction in document processing time.

SoDa Laboratories

Melbourne, VIC

AI Research Intern

Jun 2025 – Jul 2025

- Automated **60%** of enterprise data extraction workflows using **PySpark and Spark SQL** distributed processing to unblock data readiness, fine-tuning Qwen3-VL 3B via **SFT, PEFT-LoRA, and CPT**; outperformed olmOCR and DocVQA baselines on complex handwriting and layout evaluation metrics.

Foboh AI

Sydney, NSW

Machine Learning Engineer Intern

May 2025 – Jun 2025

- Co-designed ReAct multi-agent workflows on AWS Lambda for a B2B food AI startup, automating order parsing to target a **70%** ERP entry reduction; embedded with clients to establish **P95 latency** baselines via OpenTelemetry, AWS X-Ray.

Keyvalue Pencil AI

Kerala, India

Software Engineer

Mar 2023 – Jul 2024

- Enhanced conversational accuracy to **90%** and reduced retrieval latency by **40%** across **10,000+ sessions** by embedding with client teams to engineer different production RAG architectures using LangGraph, LangChain, ChromaDB, & Redis.
- Instrumented **MLflow** experiment tracking and **Model Registry** versioning across production RAG deployments, logging retrieval configurations, prompt versions, and evaluation scores to maintain reproducible benchmarks and model lineage across client environments.
- **Co-led enterprise-wide AI strategy** and implementation at KeyValue, delivering full-stack solutions for **6+ B2B clients** by bridging machine learning models with scalable web architectures and providing core technical guidance to non-technical stakeholders.
- Drove a **25%** CTR uplift across **40 brand campaigns** by collaborating directly with **Pencil AI stakeholders** to architect custom generative ad pipelines integrating Stable Diffusion, Imagen, and Adobe Firefly.

Bigbinary

Delhi, India

Software Engineer

Aug 2022 – Feb 2023

- Generated **\$250K** in revenue and increased user retention by **15%** across **21+** products by standardizing push notifications via a custom Rails Engine and **open-sourcing a reusable package** (@bigbinary/neetoui), driving a **40%** gain in operational efficiency using ReactJS, Ruby on Rails, and React Native.

TECHNICAL SKILLS

AI Governance, Evaluation & Observability: Langfuse, RAGAS, LLM-as-Judge, Harness Engineering, OpenTelemetry, Prometheus, Grafana, Fairness Evaluation, Bias Detection, Discrimination Metrics, Secure AI Guardrails, PII Obfuscation, Tenant Isolation, Role-Based Access Control (RBAC), Data Privacy (APP principles).

Generative AI & Agentic Systems: LangGraph, Agentic AI, Multi-Agent Systems, MCP Servers, RAGs, VLMs (Vision

Language Models), PEFT-LORA, Mixture of Experts (MoE), Foundation Models, OpenAI SDK, Google ADK.

Data & ML Platform: Databricks, MLflow, Model Registry, Model Serving (vLLM), Apache Spark (PySpark, Spark SQL), Parquet/Data Lakes, Distributed Inference, Multi-GPU Handling, PyTorch, TensorFlow, XGBoost, Scikit-Learn, NLP, Computer Vision, Deep Learning, Time-Series Forecasting, Reinforcement Learning.

Infrastructure & Ops: Python, SQL, C++, Java, R, AWS (SageMaker, Lambda, EC2), GCP (Cloud Run, GCS, BigQuery), Docker, Kubernetes, Terraform, FastAPI, Apache Kafka, PostgreSQL, Redis, Git

RESEARCH WORK AND PROJECTS

AI Evaluation & Observability Pipeline — *Langfuse, RAGAS, GCP (Cloud Run, GCS), LLM-as-Judge*

- Built an enterprise-grade AI observability and tracing pipeline utilizing Langfuse and GCP infrastructure to capture real-time execution graphs and JSON payload metrics across distributed model clusters.
- Benchmarked state-of-the-art Vision Language Models (VLMs) against massive handwritten archival datasets using automated LLM-as-Judge frameworks (RAGAS), reducing manual evaluation cycles by **80%**.

SignalStack — *Apache Kafka (KRaft), PostgreSQL, Redis, FastAPI, Next.js, Terraform, AWS, Prometheus, Grafana*

- Architected a self-hosted tech-trend platform ingesting real-time activity from GitHub and Hacker News, scoring trending CS topics via a custom Apache Kafka (KRaft) pipeline on a single free-tier AWS EC2 instance.
- Built a fault-tolerant consumer pipeline with partitioned Kafka topics, batched PostgreSQL writes, and Redis deduplication, raising system throughput **~250x** at zero added infrastructure cost.
- Served platform data via a **9**-endpoint FastAPI microservice with high-speed Redis caching layers and a responsive Next.js frontend; provisioned full infrastructure as code utilizing Terraform and executed real-time monitoring with Prometheus and Grafana dashboards.

Multi-Agent AI Orchestration Platform — *LangGraph, Redis, MCP Servers, Docker* **Economics Dept Project**

- Designed and deployed a LangGraph multi-agent orchestration platform for the Monash Economics Department to automate collaborative reasoning, structured schema validation, and hierarchical data processing across independent subnet agents.
- Engineered automated hierarchical delegation algorithms allowing sub-agents to replicate and synthesize empirical codebases, cutting complex econometric data synthesis cycles by **70%** while preserving exact state machine auditing.

Graduate Machine Learning Researcher — *MLLMs, PyTorch, Distributed Systems* **University of Melbourne**

- Developing fault-tolerant distributed processing pipelines under Prof. Kathryn Baragwanath to automate layout extraction and document structural understanding across **2,200+** historical archival assets.
- Implemented concurrency controls and batching primitives to ensure consistent optimization thresholds, reducing tail latency to **<7** seconds across highly unformatted document structures.

TORCH Platform — *AWS (EC2, S3, Beanstalk, IAM, VPC), MedGemma VLM, PostgreSQL* **AI Healthcare Venture**

- Targeted a **\$1B** litigation exposure reduction across **50,000+** beds/year by partnering with clinical healthcare stakeholders to engineer a highly secure Electronic Health Record (EHR) platform extension for medical handovers.
- Spearheaded serverless integration of MedGemma VLMs on AWS systems to automate legacy ISBAR handover parsing, achieving a **30%** reduction in omission errors; architected a secure human-in-the-loop validation layer utilizing AWS IAM and VPC isolation, ensuring data privacy and safety across **13+** operational scenarios.

EDUCATION

Monash University

Master's in Data Science (Machine Learning)

Melbourne, VIC

Jul 2024 – Jun 2026

- Research: Efficient MOE Architectures, Tokenizer Optimality, Multimodal Benchmarks, MLLMs, Vision, NLP

Kerala Technological University

Bachelor's in Technology

Kerala, India

2018 – 2022

- Research: Computer Vision, NLP, Autonomous Systems, IoT Systems, Business Economics

PUBLICATIONS AND AWARDS

- **Certification:** Databricks Certified Machine Learning Professional (*In Progress*).
- **Publication:** A.Joy, J. Roy, IoT Enabled Electronic Security and Remote Monitoring for Electric Bikes, IEEE IPRECON 2022.
- **Double Victory (Best Investor Pitch & People's Choice):** Won both major awards out of 21 teams for TORCH, judged by Google AI for Startups, Antler, and veteran VC partners at Monash DeepTech Entrepreneurship 2026.
- **Monash Impact Innovation Award:** Awarded 1st place at Monash ImpactHack 2025; featured on the Monash Entrepreneurship.
- **Google DSC President:** Selected among 1,300+ global leads; applied a founder's mindset, driving **3000%** membership growth.
- **Winter Research Scholarship:** Appointed to conduct Machine Learning research at SoDa Labs from amongst 10,000+ students.