

Larry Jiang

Senior AI Engineer | Generative AI | Agentic AI | RAG | MLOps

Phone: 0416 300 005

Email: weixuan.jiang@gmail.com

Melbourne, Australia

<https://www.linkedin.com/in/larryjiangai>

PROFESSIONAL SUMMARY

AI Engineer and technical architect with 5 years of experience designing, building and deploying production AI, machine learning and data systems across banking, cloud services, workplace safety and healthcare. Expertise spans Generative AI, agentic workflows, Retrieval Augmented Generation (RAG), LLM orchestration, model fine-tuning, MLOps, APIs, cloud-native deployment and data engineering. Hands-on with Python, SQL, AWS, Databricks, LangChain, LangGraph, Amazon Bedrock, SageMaker, OpenSearch, FastAPI, Docker and Kubernetes. Proven delivery of secure and observable AI services with measurable outcomes, including an 80% reduction in release validation effort and business lending assessment cycles reduced from weeks to hours.

CORE SKILLS

- Programming and Backend: Python, SQL, FastAPI, REST APIs, WebSockets, microservices, event-driven architecture, Git, Pydantic and structured outputs
- Generative and Agentic AI: LangChain, LangGraph, CrewAI, Amazon Bedrock, AWS AgentCore, Model Context Protocol (MCP), tool calling, multi-agent systems, prompt engineering and context engineering
- Machine Learning and LLMs: PyTorch, Hugging Face, SageMaker, Llama 2, Gemini, LoRA, QLoRA, PEFT, FSDP, quantisation, distributed training, inference optimisation and vLLM
- RAG and AI Evaluation: Amazon OpenSearch, embeddings, vector search, semantic search, hybrid and MMR retrieval, chunking, reranking, LLM-as-a-Judge, recall, precision, hallucination detection and human feedback
- MLOps, LLMops and Cloud: MLflow, LangSmith, LangFuse, Docker, Kubernetes, ECR, EKS, CI/CD, Jenkins, Terraform, AWS, Azure, GCP, monitoring, tracing and model lifecycle management
- Data Engineering: Databricks, PySpark, AWS Glue, S3, Redshift, PostgreSQL, Airflow, MWAA, Step Functions, Lambda and Azure Data Factory

SELECTED ENTERPRISE IMPACT

- 80% reduction in release-readiness validation effort across more than 10 service domains.
- Weeks to hours reduction in business lending assessment cycles using policy-cited agentic workflows.
- 30% improvement in retrieval accuracy through hybrid search, table-aware chunking and reranking.
- 87% reduction in operational cost for an intelligent document-processing proof of concept.

PROFESSIONAL EXPERIENCE

Commonwealth Bank of Australia, Melbourne | Aug 2025 - Sep 2026

Senior AI Engineer (Contractor via Rackspace Technology)

- Led the architecture and delivery of an enterprise business-lending covenant recommendation platform using a four-node LangGraph workflow, parallel specialist agents and policy-cited Pydantic outputs, reducing assessment cycles from weeks to hours.
- Designed a two-tier RAG architecture on Amazon OpenSearch with table-aware chunking, hybrid retrieval and reranking, eliminating hallucinated financial thresholds and enabling citations to exact policy sections; improved recommendation accuracy from 50% to 75% using LLM-as-a-Judge and analyst feedback across more than 20 industry verticals.
- Architected a release-readiness multi-agent service spanning more than 10 service domains and four business units, reducing manual validation effort by 80% while supporting two production release cycles each month.
- Engineered production agent controls for state management, retries with exponential backoff, bounded tool and model calls, structured error handling, model fallback chains and graceful degradation across three LLMs.
- Built an AIOps proof of concept using LangGraph, Amazon Bedrock, Bedrock Knowledge Bases, MWAA, Glue, S3 and Redshift to analyse logs, generate evidence-based root-cause timelines and support conversational and batch workflows.
- Designed an OAuth2 and PKCE authentication sidecar for PingID-secured MCP servers and deployed managed agent workloads through AWS AgentCore, aligning AI services with enterprise identity and platform controls.
- Implemented GenAI evaluation and observability covering retrieval recall and precision, hallucination detection, trace analysis, LLM-as-a-Judge and human feedback loops.

Amazon Web Services, Melbourne | Feb 2025 - Aug 2025

Cloud Engineer - AI/ML (Full Time)

- Acted as a technical advisor to enterprise customers on secure, scalable and cost-effective AI and ML architectures using Amazon SageMaker, AWS Bedrock and cloud-native services.
- Resolved complex production incidents across SageMaker and Bedrock, restoring availability and guiding customers on reliability, observability and operational best practice.
- Collaborated with Solutions Architects and enterprise customers to validate foundation-model applications and data pipelines through targeted proofs of concept and technical spikes.
- Designed an intelligent invoice-processing proof of concept with Bedrock and Amazon Textract that reduced manual effort by 80% and operational cost by 87%.

JR Academy, Australia (Remote) | Sep 2024 - May 2026

Generative AI Instructor (Contractor, Weekend and After-hours)

- Developed and delivered applied training for ML engineers and data scientists across RAG, semantic search, agentic systems, LLM fine-tuning, quantisation and distributed inference.

- Mentored practitioners through end-to-end projects from data ingestion and model development to evaluation and deployment, helping uplift AI engineering capability.
- Curated practical case studies that translated emerging AI methods into repeatable engineering patterns and production delivery practices.

Forwood Safety, Australia (Remote) | Apr 2024 - Oct 2024

AI Technical Architect (Full Time)

- Led cross-functional delivery across engineering, UX, subject-matter experts, DevOps and product, aligning Generative AI architecture with product outcomes and improving delivery efficiency by 20%.
- Architected a production RAG pipeline that ingested documents from S3, generated embeddings and served MMR and hybrid search through Amazon OpenSearch, increasing retrieval accuracy by 30% and reducing processing time by 40%.
- Delivered an end-to-end AI risk-analysis solution producing 80-90% expert-validated accuracy, with LangSmith evaluation and repeatable MLOps patterns using SageMaker, ECR, EKS and MLflow.
- Presented architecture, delivery risks and business value to senior stakeholders, securing alignment and supporting transparent technical decision-making.

Accenture Australia, Melbourne | Aug 2022 - Apr 2024

Senior Analyst, Data Engineer (Full Time)

- Designed Generative AI and data solutions for healthcare and financial-services clients, combining enterprise retrieval, structured prompts, multimodal processing and production observability.
- Implemented an MLOps lifecycle for fine-tuning and deploying a Llama 2 7B medical-coding model, including model versioning, containerisation, continuous performance tracking and automated evaluation for ICD-10-AM workflows.
- Built Databricks and Azure Data Factory pipelines for scalable data ingestion, transformation and integration supporting AI and analytics workloads.
- Engineered AWS ETL pipelines with Glue, PySpark, Step Functions and Lambda, achieving 99.5% data accuracy and reducing manual intervention by 95%.
- Improved retrieval relevance, precision and recall by 30% through hybrid search, reranking and prompt engineering; delivered vision-based document quality assessment at 95% accuracy.

National Australia Bank, Melbourne | Jan 2022 - Aug 2022

Analyst, Data Engineer (Full Time)

- Maintained 99% pipeline uptime across PySpark, S3, Redshift, Databricks and Airflow workloads processing streaming and transactional data at billion-row scale.
- Automated ETL validation and API ingestion, removing 20 hours of manual work per month and a further 12 hours per week from recurring operational processes.
- Optimised Redshift storage and query performance through compression and distribution design, and improved production alerting reliability.

National Australia Bank, Melbourne | Jul 2021 - Jan 2022

Intern Cloud Engineer (Full Time)

- Supported Terraform and Jenkins delivery patterns, including GitHub-triggered builds, plugin configuration, automated EC2 AMI backups and scheduled database refresh workflows.

PROJECTS

CoinBell AI, Melbourne | Sep 2025 - Present

Technical Founder

- Founded and engineered a production Generative AI cryptocurrency intelligence platform integrating real-time market data, on-chain indicators, sentiment signals and chart images for AI-assisted research.
- Built multimodal Gemini workflows and multi-agent services for market analysis, portfolio analytics and AI news, with Python APIs, structured JSON outputs, Supabase caching and credit-based model routing.
- Owned end-to-end product architecture, data ingestion, backend and frontend integration, deployment, reliability and user experience. Live product: <https://coinbell-ai.com/>

EDUCATION

Bachelor of Data Analytics | University of South Australia | Adelaide, SA | Jul 2019 - Oct 2023

CERTIFICATIONS

- AWS Certified Machine Learning - Specialty, Amazon Web Services, 2025
- AWS Certified Machine Learning Engineer - Associate, Amazon Web Services, 2025
- AWS Certified Data Engineer - Associate, Amazon Web Services, 2025
- AWS Certified AI Practitioner, Amazon Web Services, 2025
- Databricks Certified Associate Developer for Apache Spark 3.0, Databricks, 2022
- Google Cloud Certified Professional Machine Learning Engineer, Google, 2022
- Google Cloud Certified Professional Cloud Architect, Google, 2023
- Google Cloud Certified Professional Data Engineer, Google, 2022
- Google Cloud Certified Professional Cloud Developer, Google, 2023
- Kubernetes and Cloud Native Associate, The Linux Foundation, 2023
- Microsoft Certified: Azure AI Engineer Associate, Microsoft, 2021
- Microsoft Certified: Azure Data Engineer Associate, Microsoft, 2022
- Microsoft Certified: Azure Data Scientist Associate, Microsoft, 2021
- AWS Certified Cloud Practitioner, Amazon Web Services, 2021