

Larry Jiang

Phone: 0416300005

Email: weixuan.jiang@gmail.com

Melbourne, Australia

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

PROFESSIONAL SUMMARY

Senior AI Engineer with 5 years of experience building production-grade Generative AI systems across RAG, agentic workflows, MLOps, cloud engineering, and data platforms. Skilled in designing enterprise AI solutions using LangGraph, AWS Bedrock, SageMaker, OpenSearch, FastAPI, and Docker. Proven track record delivering measurable business outcomes across banking, workplace safety, and healthcare domains through AI agents, RAG pipelines, model fine-tuning, and scalable enterprise AI enablement.

SKILLS

- **Generative AI:** LangChain, LangGraph, CrewAI, AWS Bedrock, AWS AgentCore, MCP (Model Context Protocol), Pydantic, Hugging Face, Fine-Tuning (LoRA, QLoRA), Inference Optimization, Distributed Training
- **RAG & Vector Search:** Amazon Opensearch, PostgreSQL, Hybrid Search, MMR Retrieval, Embedding, Semantic Chunking, Reranking
- **MLOps & Observability:** AWS Sagemaker, MLflow, LangSmith, LangFuse, Docker, AWS ECR, AWS EKS, Airflow, AWS Step Functions, Jenkins, Terraform, AWS Cloudformation
- **Cloud & Backend:** AWS, Azure, FastAPI, WebSocket, OAuth2, Python.
- **Project Management & Leadership:** Agile Methodology, Scrum, Jira, Team Leadership, Mentoring.
- **AI Coding:** Claude Code, Codex

INDEPENDENT AI PROJECT

CoinBell AI, Melbourne

Sep 2025 - Current

Technical Founder

- Founded and engineered a GenAI-powered cryptocurrency intelligence platform integrating real-time market data, on-chain indicators, chart images, and sentiment signals for AI-assisted trading research.
- Built multi-modal AI workflows using Google Gemini to generate market summaries, interpret chart patterns, and explain potential price movements from structured and visual inputs
- Designed backend APIs and data pipeline for market data ingestion, AI inference, and analytics delivery, supporting a live production demo.
- Owned full product delivery across architecture, frontend/backend integration, deployment, multi-agent system and user experience.
- Live demo: <https://coinbell-ai.com/>

PROFESSIONAL EXPERIENCE

Commonwealth Bank of Australia, Melbourne

Aug 2025 – Sep 2026

Senior AI Engineer (Contractor via Rackspace Technology)

- Architected an agentic workflow covenant recommendation system using a 4-node LangGraph DAG with parallel Covenant and Undertaking agents fed by a shared Industry Analyst agent, reducing the business loan assessment cycles from weeks to hours with fully policy-cited, Pydantic-enforced structured outputs.
- Built a self-improving AI system where credit analyst override decisions trigger a LLM judge to extract confidence-scored patterns stored in Amazon Opensearch, retrieved as tool calls during future agent runs, compounding institutional knowledge across 20+ industry verticals and improving recommendation accuracy from 50% to 75%.
- Eliminated hallucinated financial thresholds by designing a two-tier RAG pipeline over Amazon Opensearch with table-aware chunking and re-ranking, enabling agents to cite exact policy sections rather than infer thresholds from priors.
- Architected and designed an intelligent release readiness multi-agent system using LangGraph with RAG capabilities, reducing manual validation time by 80% across release management, developer, and platform engineering teams.
- Implemented robust agent orchestration—including state management, error handling, and automated retries—using LangGraph's graph-based execution to support two monthly releases across 10+ service domains.
- Built a conversational release-readiness agent with advanced prompt engineering, context-aware reasoning, and multi-turn dialogue, leveraging context engineering to synthesize evidence from multiple sources and support queries across four federated business units.
- Designed and deployed a SSO authentication sidecar container for implementing OAuth2 + PKCE with PingID to securely manage authentication flows for MCP (Model Context Protocol) servers.
- Engineered production agent reliability using middleware stack (tool call limit, model call limit, model fallback chain, retry with backoff) via custom agent framework, achieving graceful degradation across primary and two fallback LLMs
- Deployed the AI agent using AWS AgentCore, utilizing managed orchestration and secure service integration to deliver scalable, production-grade agent operations.
- Experienced in building evaluation frameworks for GenAI systems, including LLM-as-a-Judge workflows, retrieval quality assessment (recall and precision), hallucination detection, trace analysis, and human feedback loops to improve production reliability.

Amazon Web Services, Melbourne

Feb 2025 – Aug 2025

Cloud Engineer – AI/ML (Full Time)

- Resolved complex production incidents across Amazon SageMaker and AWS Bedrock, ensuring high availability for AI/ML workloads and minimising downtime.
- Advised enterprise customers on best practices for scalable, secure and cost-effective AI/ML deployments in AWS leading to improved operational efficiency.
- Collaborated with Solutions Architects and enterprise clients to design PoCs leveraging foundation models and pipelines for generative AI.

Key Achievement(s):

- Designed and implemented a Proof of Concept (PoC) for intelligent invoice processing using AWS Bedrock and Amazon Textract, achieving an 80% reduction in manual effort and an 87% decrease in operational costs through automation and Generative AI.
- Created a restaurant recommendation engine prototype integrating generative AI and user preference data to dynamically rank and personalize menu suggestions, showcasing expertise in recommendation systems.

JR Academy, Australia (Remote)

Sep 2024 – May 2026

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

- Developed and delivered online lectures and practical sessions on generative AI, LLM fine-tuning, RAG, semantic search and agent-based systems to ML engineers and data scientists.
- Delivered advanced training on distributed fine-tuning (Lora, QLora) and quantization methods, referencing frameworks such as Hugging Face PEFT, PyTorch FSDP and vLLM.
- Mentored professionals through hands-on projects, guiding them in building AI solutions from data ingestion to deployment.
- Curated case studies to translate theoretical knowledge into practical applications, fostering a strong learning culture.

Forwood Safety, Australia (Remote)

April 2024 – Oct 2024

AI Technical Architect (Full Time)

- Led cross-functional teams (web, UX) and coordinated with SMEs, DevOps, and product managers, aligning Generative AI initiatives with business goals and improving project efficiency by 20%.
- Architected and implemented automated data pipelines to preprocess documents from S3, generate embeddings (OpenAI), and integrate with Amazon OpenSearch for MMR/hybrid search, boosting retrieval accuracy by 30% and reducing processing time by 40%.
- Designed and implemented an end-to-end Generative AI solution achieving 80–90% accuracy in expert-validated insights, enhancing risk analysis and decision-making.
- Developed and refined AI Agents for risk analysis, integrating LangSmith for evaluation and improving generative accuracy and operational efficiency.
- Developed scalable MLOps pipelines using AWS SageMaker, ECR, EKS, and MLflow, enabling comprehensive model versioning, performance tracking, and AI solution reproducibility.
- Presented technical solutions to senior stakeholders, promoting transparency, securing buy-in, and streamlining Agile delivery processes.

Key Achievement(s):

- Architected and implemented a RAG solution that improved the retrieval accuracy by 30% and reduced processing time by 40%

Accenture Australia, Melbourne

Aug 2022 – April 2024

Senior Analyst, Data Engineer (Full Time)

- Architected and optimized Generative AI-driven healthcare systems, improving data retrieval accuracy by 30% and enhancing reliability through real-time observability, reducing downtime by 25%.
- Designed specialized prompt engineering and Retrieval Augmented Generation (RAG) techniques—including re-ranking and hybrid search—boosting relevance, precision, and recall by 30%.
- Implemented comprehensive MLOps pipeline for fine tuning Llama 2 (7B) medical coding model, featuring end-to-end model versioning, containerized deployment, continuous performance tracking, and automated evaluation pipelines that significantly elevated ICD-10-AM coding accuracy for healthcare applications.
- Developed structured prompts and vision-based captioning systems for financial clients, achieving 95% accuracy in quality assessments and cutting document retrieval times by 40%.
- Designed and implemented scalable data pipelines using Databricks and Azure Data Factory, enabling efficient data ingestion, transformation, and integration for ICD-10-AM coding.
- Engineered AWS-based data pipelines with ETL (AWS Glue, PySpark) ensuring 99.5% data accuracy, while automating infrastructure tasks (AWS Step Functions) to reduce manual interventions by 95%.
- Supported Auto ML prototypes and streamlined ETL triggers (AWS Lambda), ensuring timely, clean data ingestion and faster model experimentation.

Key Achievement(s):

- Drove stakeholder engagement and secured funding by effectively communicating technical insights and business value.
- Implemented cutting-edge prompt engineering and model fine-tuning techniques to deliver measurable performance gains.

National Australia Bank, Melbourne

Jan 2022 – Aug 2022

Analyst, Data Engineer (Full Time)

- Ensured 99% pipeline uptime by efficiently ingesting streaming and transactional data (PySpark, S3, Redshift, Python, Databricks).
- Maintained and optimized Airflow DAGs, reducing data processing disruptions and improving operational efficiency.
- Automated ETL frameworks across billions of rows, cutting 20 hours of manual work per month.
- Optimized Redshift tables via compression and distribution strategies, improving query performance and storage efficiency.
- Developed Python-based ETL pipelines for Khoros API, saving 12 hours of manual work weekly.
- Collaborated in Agile teams with data engineers, product owners, data scientists, and software engineers.

Key Achievement(s):

- Automated data validation between databases with Python and Airflow, ensuring accurate and error-free ingestion.
- Resolved Airflow email server issues in production, improving reliability and timely alerting.

National Australia Bank, Melbourne

July 2021 – Jan 2022

Intern Cloud Engineer (Full Time)

- Assisted with updating Terraform template and variables.

- Configured Jenkins for installing necessary plugins.
- Configured source code management for Jenkins with GitHub build trigger.
- Implemented daily automatic backup for AMI of EC2 instances in Jenkins.
- Implemented weekly database refresh and backup in Jenkins via jenkinsfile.

EDUCATION

July 2019 – Oct 2023	University of South Australia	Adelaide, SA
	<i>Bachelor of Data Analytics</i>	

CERTIFICATIONS

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

REFEREES

-
- Available upon request