

Weixuan (Larry) Jiang

Melbourne, Australia | Australian Citizen

Phone: 0416300005 | Email: weixuan.jiang@gmail.com

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

PROFESSIONAL SUMMARY

AI Technical Architect and Senior AI Engineer with 5 years of experience across enterprise AI, cloud engineering and data platforms. Hands-on experience designing and implementing AI agents, secure service integrations and production operations in banking, healthcare and workplace safety. Combines architecture and delivery using LangGraph, AWS Bedrock, AgentCore, SageMaker and OpenSearch, with a focus on reliable AI adoption, stakeholder alignment and measurable business outcomes.

SKILLS

- **AI Architecture:** End-to-end GenAI solution design, multi-agent orchestration, PoC development, production deployment, stakeholder engagement and cross-functional delivery.
- **Generative AI:** LangGraph, LangChain, CrewAI, AWS Bedrock, AWS AgentCore, MCP, Pydantic, Hugging Face, LoRA and QLoRA.
- **Security and Reliability:** OAuth2, PKCE, PingID SSO, authentication sidecars, agent state management, retries, model fallback, LLM-as-a-Judge, human feedback and trace analysis.
- **RAG and Search:** Amazon OpenSearch, PostgreSQL, hybrid search, MMR retrieval, embeddings, semantic and table-aware chunking, reranking.
- **Cloud and MLOps:** AWS, Azure, SageMaker, Docker, ECR, EKS, MLflow, LangSmith, LangFuse, Terraform, CloudFormation, Jenkins, Airflow and Step Functions.
- **Engineering and Delivery:** Python, FastAPI, WebSocket, Databricks, Azure Data Factory, AWS Glue and PySpark; Agile, Scrum, Jira, mentoring; Claude Code and Codex.

PROFESSIONAL EXPERIENCE

Commonwealth Bank of Australia, Melbourne

Aug 2025 - Sep 2026

Senior AI Engineer (Contractor via Rackspace Technology)

- Architected a covenant recommendation system using a four-node LangGraph DAG, with parallel Covenant and Undertaking agents supported by a shared Industry Analyst agent. Reduced business loan assessment cycles from weeks to hours through policy-cited, Pydantic-enforced structured outputs.
- Designed a two-tier RAG pipeline over Amazon OpenSearch with table-aware chunking and reranking, eliminating hallucinated financial thresholds by grounding agent responses in exact policy sections.
- Designed and deployed an SSO authentication sidecar using OAuth2 and PKCE with PingID to secure authentication flows for MCP servers.
- Deployed an AI agent using AWS AgentCore, integrating managed orchestration and secure services for scalable, production-grade operations.
- Engineered production reliability through tool and model call limits, retry with backoff and a model fallback chain, enabling graceful degradation across a primary and two fallback LLMs.
- Architected a release-readiness multi-agent system using LangGraph and RAG, reducing manual validation time by 80% across release management, developer and platform engineering teams.
- Implemented state management, error handling and automated retries to support two monthly releases across 10+ service domains. Built conversational evidence synthesis and multi-turn queries across four federated business units.
- Built a human-feedback loop in which credit analyst overrides triggered an LLM judge to extract confidence-scored patterns for reuse through OpenSearch tools across 20+ industry verticals, improving recommendation accuracy from 50% to 75%.
- Experienced in building GenAI evaluation frameworks covering LLM-as-a-Judge, retrieval precision and recall, hallucination detection, trace analysis and human feedback to improve production reliability.

PROFESSIONAL EXPERIENCE CONTINUED

Amazon Web Services, Melbourne

Feb 2025 - Aug 2025

Cloud Engineer - AI/ML (Full Time)

- Advised enterprise customers on scalable, secure and cost-effective AI/ML deployments in AWS, and collaborated with Solutions Architects and clients to design generative AI proofs of concept.
- Resolved complex production incidents across Amazon SageMaker and AWS Bedrock, supporting high availability and minimising downtime for AI/ML workloads.
- Designed and implemented an intelligent invoice-processing PoC using AWS Bedrock and Amazon Textract, achieving an 80% reduction in manual effort and an 87% decrease in operational costs.
- Created a generative AI restaurant recommendation prototype combining user preferences with dynamic menu ranking and personalisation.

JR Academy, Australia (Remote)

Sep 2024 - May 2026

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

- Developed and delivered lectures and practical sessions on generative AI, LLM fine-tuning, RAG, semantic search and agent-based systems for ML engineers and data scientists.
- Delivered advanced training in LoRA, QLoRA, distributed fine-tuning and quantisation, referencing Hugging Face PEFT, PyTorch FSDP and vLLM.
- Mentored professionals through hands-on AI projects from data ingestion to deployment, using case studies to connect theory with practical delivery.

Forwood Safety, Australia (Remote)

Apr 2024 - Oct 2024

AI Technical Architect (Full Time)

- Led cross-functional web and UX teams and coordinated with SMEs, DevOps and product managers to align GenAI initiatives with business goals, improving project efficiency by 20%.
- Architected and implemented document pipelines from S3 through OpenAI embeddings to Amazon OpenSearch for MMR and hybrid search, improving retrieval accuracy by 30% and reducing processing time by 40%.
- Designed and implemented an end-to-end GenAI solution delivering 80-90% accuracy in expert-validated insights for risk analysis and decision-making.
- Developed AI agents for risk analysis and integrated LangSmith evaluation to improve generative accuracy and operational efficiency.
- Built scalable MLOps pipelines using SageMaker, ECR, EKS and MLflow for model versioning, performance tracking and reproducible AI delivery.
- Presented technical solutions to senior stakeholders, secured buy-in and supported Agile delivery.

Accenture Australia, Melbourne

Aug 2022 - Apr 2024

Senior Analyst, Data Engineer (Full Time)

- Architected and optimised GenAI healthcare systems, improving data retrieval accuracy by 30% and reducing downtime by 25% through real-time observability.
- Designed RAG and prompt engineering techniques, including reranking and hybrid search, to improve relevance, precision and recall by 30%.
- Implemented an end-to-end MLOps pipeline for fine-tuning a Llama 2 7B medical coding model, covering versioning, containerised deployment, performance tracking and automated evaluation for ICD-10-AM coding.
- Built Databricks and Azure Data Factory pipelines for ICD-10-AM data ingestion, transformation and integration. Engineered AWS Glue and PySpark ETL with 99.5% data accuracy and used Step Functions to reduce manual intervention by 95%.
- Developed structured prompts and vision-based captioning for financial clients, achieving 95% quality-assessment accuracy and cutting document retrieval times by 40%.
- Supported AutoML prototypes and Lambda ETL triggers, and communicated technical insights and business value to engage stakeholders and secure funding.

PROFESSIONAL EXPERIENCE CONTINUED

National Australia Bank, Melbourne

Jan 2022 - Aug 2022

Analyst, Data Engineer (Full Time)

- Delivered streaming and transactional data ingestion using PySpark, S3, Redshift, Python and Databricks, ensuring 99% pipeline uptime.
- Maintained and optimised Airflow DAGs and automated ETL across billions of rows, saving 20 hours of manual work per month.
- Optimised Redshift compression and distribution strategies. Built Python ETL pipelines for the Khoros API, saving 12 hours of manual work weekly.
- Automated cross-database validation with Python and Airflow, resolved production alerting issues and collaborated with product owners, data scientists and software engineers in Agile teams.

National Australia Bank, Melbourne

Jul 2021 - Jan 2022

Intern Cloud Engineer (Full Time)

- Assisted with updating Terraform templates and variables, and configured Jenkins plugins and GitHub build triggers.
- Implemented daily EC2 AMI backups and weekly database refresh and backup workflows through Jenkins.

INDEPENDENT AI PROJECT

CoinBell AI, Melbourne

Sep 2025 - Current

Technical Founder

- Founded and engineered a GenAI cryptocurrency intelligence platform integrating real-time market data, on-chain indicators, chart images and sentiment for AI-assisted trading research.
- Owned architecture, backend APIs, ingestion and inference pipelines, frontend/backend integration, multi-agent systems and deployment, delivering a live production demo.
- Built multimodal Google Gemini workflows to summarise market data, interpret charts and explain potential price movements from structured and visual inputs.
- Live demo: <https://coinbell-ai.com/>

EDUCATION

University of South Australia, Adelaide, SA

Jul 2019 - Oct 2023

Bachelor of Data Analytics

SELECTED 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 Cloud Architect, Google, 2023
- Google Cloud Certified Professional Cloud Developer, Google, 2023
- Google Cloud Certified Professional Database Engineer, Google, 2023
- Kubernetes and Cloud Native Associate, The Linux Foundation, 2023
- Google Cloud Certified Professional Machine Learning Engineer, Google, 2022
- Google Cloud Certified Professional Data Engineer, Google, 2022
- Microsoft Certified: Azure Data Engineer Associate, Microsoft, 2022
- Databricks Certified Associate Developer for Apache Spark 3.0, Databricks, 2022

REFEREES

Available upon request