

Delai Ye

AI Engineer | LangGraph & Agent Orchestration & RAG | Multi-Agent Systems · Full-Stack

Melbourne, VIC | Australian PR | Chinese Citizen | alei198634@gmail.com | 0424235896

[Portfolio](#) | [GitHub](#) | [LinkedIn](#)

SUMMARY

- Currently AI Engineer in Clenergy, co-building an **Enterprise-Level AI platform** — a tiered knowledge platform routing each question to curated exact Q&A, cited document RAG, or governed Text2SQL on a self-built orchestrator with 100% (20/20) end-to-end accuracy.
- Built **ASX Mining Financial Chatbot** — a multi-agent RAG chatbot covering 5 ASX mining companies × 3 fiscal years, deployed on AWS EC2 + Vercel with full CI/CD.
- Built **Research Assistant** (nested graphs, human-in-the-loop, parallel Send nodes, dual retrieval).
- Strong with Gen AI / LLMs— Text-to-SQL, RAG, Agents, harness and evaluation—and known for clear communication and high execution speed from prototype to production.
- Highly interested in automating day-to-day workflows by combining scripts and/or AI agents (FDE-style).

AI AGENT & LLM ENGINEERING SKILLS

- **Agent frameworks** like LangGraph, LangChain, ReAct, Graph Tests & Evals, LangSmith tracing.
- **Agent patterns** Multi-agent orchestration, ReAct-style tool use, long-horizon tasks, ambient/event-driven agents, subagent coordination.
- **RAG and tools** Hybrid dense + BM25 retrieval, ChromaDB vector store, AI embeddings, BM25 keyword index, semantic cache, Tavily web search, prompt engineering, structured output synthesis.
- **LLM engineering** Pydantic structured output, TypedDict state schemas, prompt templates, output parsers, streaming.
- **MCP and agent tooling** Model Context Protocol concepts, MCP servers/resources/tools, Claude Code, Cursor, ChatGPT, agentic debugging and multi-file code edits.

FULL-STACK SKILLS

- **Backend / AI app** Python, FastAPI, REST APIs, Server-Sent Events, Pydantic, SQLite, PostgreSQL basics, Docker, Uvicorn.
- **Frontend** React, Vite, Next.js, JavaScript, TypeScript, HTML, CSS, Tailwind CSS, React Query, Zustand, GSAP.
- **Cloud / DevOps** AWS (Docker production deployment, GitHub Actions CI/CD), Railway, Vercel, Cloudflare Workers, D1/R2, Linux, Nginx, SSL, PM2, CI/CD, Git.
- **Web3** Solidity, Foundry, ethers.js, Chainlink VRF, OpenZeppelin ERC-721, smart contract testing and deployment.

FEATURED AI PROJECTS

Enterprise-Level AI platform | Tiered enterprise knowledge & analytics agent platform

Python | FastAPI | Pydantic | SQLAlchemy | React | TypeScript | PostgreSQL + pgvector | MySQL | OpenAI GPT-5 | cross-encoder rerank | MinerU | Docker | AWS EC2

- Tiered knowledge platform routing every question to curated exact Q&A, cited document RAG, or constrained Text2SQL — on a self-built orchestrator rather than LangChain, with knowledge retrievable only after human review.
- Document RAG: MinerU parsing → heading-aware chunking → hybrid vector + full-text retrieval → RRF fusion → cross-encoder reranking, with context capped at top-5 (+14pp over fusion alone).
- Governed Text2SQL where the model never emits SQL, only adjusting filters or columns inside human-reviewed templates; 100% (20/20) end-to-end accuracy.

ASX Mining Financial Chatbot | Production-level LangGraph-based multi-agent RAG chatbot | [Link](#) | [github](#)

Python | LangGraph | LangChain | FastAPI | Next.js | PostgreSQL | ChromaDB | BM25 | Docker | AWS | GitHub Actions | Vercel

- LangGraph supervisor graph orchestrating three agents in parallel: hybrid retrieval (Jina dense + BM25, fan-out per company with auto-retry grading), a Tavily news agent, and a ReAct calculator.
- Vision-aware PDF ingest (PyMuPDF filtering + Gemini Vision table extraction), two-tier semantic cache, SSE streaming to Next.js, deployed on AWS EC2 via GitHub Actions CI/CD.

LangGraph Research Assistant | Full-stack multi-agent research system | [Link](#) | [github](#)

Python | LangGraph | LangChain | FastAPI | React + Vite | Docker | Railway | Vercel | DeepSeek | Tavily | Exa | SQLite | Pydantic

- Nested LangGraph architecture — outer research orchestration graph plus reusable per-analyst interview subgraphs — with human-in-the-loop approval via `interrupt_before`.
- `Send()` fan-out of N parallel interviews merged through reducer state, dual retrieval (Tavily + Exa), FastAPI backend with SQLite checkpointing and SSE streaming.

PERSONAL FULL-STACK PROJECTS

Gula Stories | Cross-platform children's story app | [Link](#)

React Native | *iOS* | *Android* | *audio content* | *offline download*

- Ad-free bedtime story app shipped solo to both App Store and Google Play with age-group filtering, multiple narrators, vivid avatars, and offline download.
- Built a highly automated AI-powered children's storybook creation workflow. It covers the entire process with AI involved in multi-steps: generating and refining stories, creating image prompts from the story content, automatically controlling the browser to generate illustrations, saving the completed artwork locally, adding watermarks, and finally uploading the finished storybooks to the cloud and the app.

FamHive | Full-stack SaaS family management platform | [Link](#)

Next.js | *React* | *Cloudflare Workers* | *Hono* | *D1/R2* | *Stripe* | *Google OAuth* | *JWT*

- End-to-end serverless SaaS covering family scheduling, children's growth tracking, rewards, bill management, auth, and tiered Stripe billing.

Wheel on Chain | Blockchain lucky wheel DApp on Sepolia | [Link](#)

Next.js | *Node.js* | *WebSocket* | *MySQL* | *Solidity* | *Chainlink VRF V2+* | *ethers.js* | *Foundry*

- Live DApp with provably fair randomness, ERC-721 NFT minting, Merkle whitelist, reentrancy protection, real-time WebSocket state, and Foundry tests.

WORKING EXPERIENCE

AI Engineer | July 2025 – Present

- Co-architected an enterprise-level AI platform with “digital employee” -- precise QA and Text-to-SQL as primary answering paths with RAG as backup, plus Function-Calling / workflow agents for an “ask → query → execute” loop -- packaged as a reusable cross-department pattern.
- Partnered with the design department to convert a fully manual SolidWorks-based design workflow into a highly automated, AI-led pipeline built on Claude Code -- AI now produces the design work and engineers act mainly as reviewers, cutting end-to-end turnaround from 5 days to 1, an 80% reduction.
- Worked with the logistics team to map one of their daily workflows, then built them a custom app that automates it end to end -- cutting each task from 20 minutes to 1.

Solo Trader — IT Developer | Independent | 2025 – Present

- Operate an independent software development practice focused on building consumer-facing mobile applications.

Independent Investor | Self-directed | 2020 – Present

- Manage a personal investment portfolio; developed strong analytical and independent decision-making discipline.

Research Scientist | VSPC Ltd, Brisbane | 2018 – 2020

- Applied electrochemical and materials R&D in an industry setting; contributed to product development and testing.

Postdoc Research Fellow | The University of Queensland, Brisbane | 2014 – 2017

- Conducted materials science research; designed experiments, analysed data, and co-authored publications.

EDUCATION & AI TRAINING

ICT30120 - Certificate III in Information Technology | Swinburne University TAFE, Hawthorn | Mar 2026 - Jun 2026

- Networking, IT support, cyber security, cloud computing, programming, and web development.

PhD – Chemical Engineering | University of Queensland, Brisbane | Oct 2010 - Jun 2014

- Thesis: Development of High Energy Cathode Materials for Lithium Ion Batteries.

Bachelor – Applied Chemistry | Central South University, China | Sep 2004 - Jun 2008

- Thesis title: The design of an Aqueous Lithium-Ion Battery with an cathode of $\text{LiNi}_{1/3}\text{Mn}_{1/3}\text{Co}_{1/3}\text{O}_2$.

LangChain Academy | 2026

- Introduction to LangChain, Introduction to LangGraph, LangSmith Essentials, Deep Agents with LangGraph

Anthropic Courses | 2026

- Claude API, Claude Code, Model Context Protocol, agent skills, subagents, prompt engineering, tool use, streaming, and production AI patterns.

Other Training

- Full-Stack Web Development Bootcamp (Udemy, 2025); W3Schools JavaScript foundations; Cyfrin/WTF/OpenZeppelin blockchain training