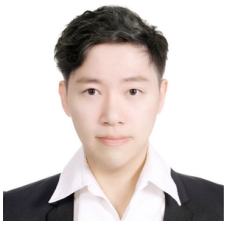

KEVIN LUO (Luo ZihJia)

AI Product Manager | Full-Stack AI Engineer | Enterprise AI Consultant

Taipei, Taiwan | kilong31442@gmail.com | github.com/kevin801221

PROFESSIONAL SUMMARY

Background

- Cross-domain AI Product Manager & Full-Stack Engineer with 5+ years spanning product management, AI engineering, medical AI, and semiconductor manufacturing
- Transitioned from nuclear physics to AI — a technical-background PM who speaks both the language of engineers and business stakeholders

Product Track Record

- Built 3 complete AI products from 0→1 as sole developer & product owner:
  - LegalSign AI — RAG-based legal assistant (LangGraph + CopilotKit)
  - FormosaEagle AI — Multi-agent smart tourism platform (Google ADK + Flutter + VIVE Eagle)
  - Casper AI — E-commerce product image designer (React + Gemini multimodal)
- Contributed as core team member on 2 enterprise-grade platforms:
  - Medical VLM Pipeline — Reinforcement learning + multimodal medical imaging
  - ESG RAG System — 7-agent ESG report generation & deep research platform
- Independently managed full business cycle: client negotiation → design → development → delivery → payment closure

PM & Leadership

- UC Berkeley AgentX Program (Jul 2025, online) — team resource allocation, cross-functional coordination, placing right people in right roles
- Reinforced PM leadership skills on top of strong technical foundation

Consulting & Recognition

- Paid AI consulting: Deloitte (automation process), Nike & Hotai Motor (Legal AI & product advisory)
- Two-time national AI competition champion | 2 papers at UC Berkeley AgentX
- Taiwan AI Academy Executive Program graduate (Class 9, 300+ executives)

TECHNICAL STACK

AI / ML: LangGraph, RAG, Multi-Agent, Google ADK, A2A, MCP, Unsloth, TRL (SFT/GRPO), VLM Fine-tuning, LoRA/QLoRA, RLHF, DPO, GRPO, Gemini 2.5/3, OpenAI, XGBoost, CV, Context Engineering

Backend: Python, FastAPI, Node.js, LangChain, LlamaIndex, WebSocket, WebRTC, GCP (Cloud Run, Vertex AI), AWS (App Runner, ECR), Docker, CI/CD

Frontend: React 18, Next.js 16, Flutter, CopilotKit, Tailwind CSS, Framer Motion, Vite

Databases: Qdrant (Vector DB), PostgreSQL, MongoDB, Redis, SQLAlchemy, SQLite
Product: Figma, Jira, Notion, Mermaid, User Research, A/B Testing, POC Validation
Automation: n8n, MCP, CI/CD, Firecrawl, OpenTelemetry

KEY STRENGTHS

Client-Facing & Revenue
Consulted for Nike, Hotai Motor, Deloitte. Independently negotiated, delivered, and closed payments.

0→1 & Team Builder
3 AI products solo + 2 enterprise platforms as core team. Ships and monetizes.

Enterprise AI Speaker
Trained at Deloitte, HSIP, top universities. Bridges AI and business for executives.

Technical PM Leadership
Technical background = engineer credibility. AgentX = resource allocation & leadership.

Cross-Functional Coordination
Bridge between product, engineering, design, business. Ambiguity to delivery.

Full-Stack Architecture
React/Flutter, FastAPI, Vector DB, LLM orchestration, Docker/AWS deployment.

ENTERPRISE CONSULTING

Deloitte Taiwan | AI Automation Process Consultant

Jan – Jun 2025

- Independently secured consulting engagement
- Advised on integrating AI-driven workflow automation into enterprise operations
- Designed corporate training on Enterprise AI Applications and No-code Agent development
- Provided AI tool selection strategy and implementation roadmaps

Nike | Legal AI Consultant & Product Advisor

2025

- Advised on AI-augmented contract review, compliance checking, and legal document management
- Assessed existing legal operations and proposed enterprise-scale AI solutions

Hotai Motor (和泰汽車) | Legal AI Consultant & Product Advisor

2025

- Evaluated AI-powered contract review for automotive procurement and dealer agreements
- Proposed phased AI adoption strategy with system integration plan

PRODUCT & TECHNICAL EXPERIENCE

LegalSign AI (律果科技) — AI Legal Assistant Platform | Technical Lead & Product Owner

Mar 2025 – Present | Taipei

- Identified market gap: Taiwan's 1.67M SMEs (95% without legal staff)
- Validated through user research, customer interviews, and competitive analysis
- Architected entire platform solo:
  - Frontend: Next.js 16 + React 19 + CopilotKit UI

  - Backend: FastAPI + LangGraph orchestration
  - Storage: Qdrant vector DB + MongoDB + MySQL
- Designed 5+ specialized LangGraph workflow agents:
  - Legal consultation (RAG) | Contract review (risk scoring)
  - Contract drafting | Contract comparison | Legal translation
- Engineered Playbook version control — competitive moat for client retention
- Deployed on GCP Cloud Run with CI/CD, 99.5% uptime
- Independently managed full business cycle: negotiation → delivery → payment

Tech: Next.js, React, CopilotKit, FastAPI, LangGraph, LangChain, Gemini 3, OpenAI, Qdrant, MongoDB, GCP, Docker, CI/CD

▸ **System Case Study**

1. What system did you build?

AI legal assistant SaaS with 5 LangGraph workflow agents serving Taiwan's SME market:

- RAG-based legal consultation with case law retrieval
- Contract review with automated risk scoring
- Contract drafting, comparison, and legal translation modules

2. How does it work?

- User uploads contracts or asks legal questions
- Supervisor agent routes to specialized sub-agents
- Each agent executes LangGraph workflow: RAG retrieval → analysis → response generation
- Results aggregated with risk scores and actionable recommendations

3. What problem did you solve?

95% of Taiwan's 1.67M SMEs lack dedicated legal staff. Traditional consultation costs NT$5,000–50,000/session. Platform reduces cost by 90%+ and delivers results in minutes.

4. Key mechanisms?

- LangGraph state machine for multi-step legal reasoning
- Adaptive RAG with Qdrant vector search for case law retrieval
- Playbook version control (Git-like) for enterprise legal rules — creating switching costs
- CopilotKit for real-time human-AI collaboration UI

5. How do you ensure stability?

- GCP Cloud Run auto-scaling with 99.5% uptime
- Gemini 3 primary + OpenAI fallback for LLM redundancy
- PostgreSQL checkpoint for conversation persistence
- CI/CD pipeline with automated testing + SonarCloud monitoring

Medical VLM — Full Training Pipeline | AI Engineer (Team Collaboration) | LegalSign AI

Jun 2025 – Present | Taipei

- Collaborated with team to build complete multimodal medical VLM fine-tuning pipeline
- Pipeline stages:
  - Data preparation (MIMIC-CXR, PathVQA, DermaBench)
  - SFT supervised fine-tuning (Unsloth + TRL SFTTrainer)
  - GRPO reinforcement learning with custom reward functions
  - Model deployment with LoRA adapter merging
- Implemented Unsloth framework: 2x training speed, 70% memory reduction
- Designed GRPO reward functions: format compliance + clinical accuracy + CoT quality scoring

Tech: Unsloth, TRL, PyTorch, LoRA/QLoRA, Vision Transformer, RLHF, DPO, GRPO, Hugging Face

▸ **System Case Study**

1. What system did you build?

End-to-end medical VLM training pipeline for multimodal medical image understanding:

- Raw medical images + clinical text → fine-tuned VLM
- Supports radiology, pathology, and dermatology modalities
- Outputs: image VQA, automated report generation, classification

2. How does it work?

- Stage 1: Curate medical datasets into standardized conversation format
- Stage 2: SFT fine-tuning on base VLM (Qwen2.5-VL / Llama 3.2 Vision / Gemma 3)
- Stage 3: GRPO RL training with custom reward functions for medical reasoning
- Stage 4: Merge LoRA adapters → deploy production model

3. What problem did you solve?

Generic VLMs hallucinate on medical images. Fine-tuned models achieve clinical-grade accuracy for specific modalities, enabling reliable automated reporting.

4. Key mechanisms?

- Unsloth weight-sharing: 90% VRAM reduction for RL training
- LoRA/QLoRA selective layer fine-tuning (vision vs. language layers)
- GRPO: no value model needed, more efficient than PPO
- Custom reward scoring: format + medical accuracy + Chain-of-Thought quality

5. How do you ensure stability?

- Multi-model support (Qwen2.5-VL, Llama 3.2, Gemma 3) as fallback
- Automatic Base64 vs. Binary image format detection
- bf16 mixed precision for numerical stability
- Medical compliance data augmentation and validation protocols

ESG RAG System — AI Smart Analysis Platform | AI Engineer (Team Collaboration) | LaplaceAI Co., Ltd.

Jun 2025 – Present | Taipei

- Built enterprise-grade ESG report generation platform
- Architecture: LangGraph Supervisor-Orchestrator with 7 specialized agents:
  - Research | Content | Compliance | Quality | Review | Finish | ReAct
- Dynamic multi-expert collaboration: 1–8 experts running in parallel
- Hybrid RAG: Qdrant vector + BM25 keyword + multi-strategy reranker
- Integrated 8 LLM providers with automatic fallback chain
- Multi-format document processing: PDF (3-engine), DOCX/PPTX/XLSX, OCR, audio
- Deployed via GitHub Actions → Docker → AWS ECR → AWS App Runner

Tech: FastAPI, LangGraph, LangChain, LlamaIndex, Qdrant, PostgreSQL, MongoDB Atlas, Docker, AWS, OpenTelemetry, A2A

▸ **System Case Study**

1. What system did you build?

Enterprise-grade ESG reporting AI platform with 7 specialized agents + deep research + hybrid RAG.

2. How does it work?

- User initiates ESG report request
- Supervisor routes to specialized agents
- Research agent: deep search (Tavily / SerpAPI / DuckDuckGo)
- Content agent: drafts sections | Compliance agent: checks regulations
- Quality agent: scores output | Review agent: refines
- Finish agent: assembles final report
- Multi-expert mode: 1–8 experts in parallel for consensus

3. What problem did you solve?

ESG reporting is mandatory but extremely labor-intensive (weeks of work). This platform automates research, drafting, and compliance — reducing production from weeks to hours.

4. Key mechanisms?

- LangGraph Supervisor-Orchestrator for complex multi-step workflows
- A2A (Agent-to-Agent) protocol for inter-agent communication
- Hybrid retrieval: Qdrant semantic + BM25 keyword + reranker
- PostgreSQL checkpoint for crash recovery & resume
- Multi-LLM factory: 8 providers with automatic fallback

5. How do you ensure stability?

- PostgreSQL checkpoint: resume-from-interruption on any failure
- 8 LLM fallback chain: Gemini → Azure OpenAI → Claude → Grok → Ollama
- OpenTelemetry tracing: Google Cloud Trace + Zipkin monitoring
- AWS App Runner auto-scaling + Docker multi-stage reproducible builds
- CI/CD with pytest automated testing

UC Berkeley AgentX Program | Researcher & Team Coordinator

Jul 2025 | Remote (US-based program)

- Intensive online research and competition program on cutting-edge AI agent systems
- Key PM takeaways:
  - Team resource allocation and cross-functional coordination
  - Assessing team members' strengths; placing right people in right roles
  - Acquired latest agent technologies directly applicable to production
- Published 2 research papers on Edge AI and DB-RAG agent systems
- Validated LegalSign AI product concept in international market

FormosaEagle AI × HTC VIVE Eagle | Product Planning & System Architect

Nov 2025 – Present | Taipei

- Designed 5 parallel AI Agent system using Google ADK Supervisor + Sub-Agent orchestration:
  - Flight | Hotel | Food | Transport | Experience
- Developed 18 API endpoints (FastAPI) + 12-screen Flutter app
- Real-time voice: OpenAI Realtime API + WebRTC
- Built 10,000+ tourism data entries in Qdrant vector DB
- Designed ESG gamification (carbon footprint tracking + sustainable tourism rewards)
- Advanced to VIVE Eagle Developer Program semifinals as sole individual participant

Tech: Flutter, FastAPI, Google ADK, A2A, MCP, OpenAI Realtime API, WebRTC, Qdrant, MongoDB, Redis, Gemini 2.5, GCP

▸ **System Case Study**

1. What system did you build?

Multi-agent smart tourism platform with 5 parallel AI agents + Supervisor, designed for HTC VIVE Eagle AR glasses with hands-free voice.

2. How does it work?

- User speaks travel request via AR glasses
- OpenAI Realtime API transcribes voice
- Supervisor decomposes into sub-tasks
- 5 agents execute in parallel (flight, hotel, food, transport, experience)

- Results aggregated → AR overlay with GPS-triggered navigation

3. What problem did you solve?

Tourists juggle 5+ apps. This unifies all travel services into one AI voice interface with AR hands-free interaction.

4. Key mechanisms?

- Google ADK Supervisor + Sub-Agent parallel orchestration
- WebRTC low-latency real-time voice
- Qdrant semantic search over 10,000+ tourism entries
- UCP (Universal Checkout Protocol) for unified payments
- Geofence-triggered GPS navigation
- SDK abstraction layer (5 interfaces) for VIVE Eagle hardware

5. How do you ensure stability?

- 5 agents run independently — single failure doesn't block others
- Redis caching for frequently accessed data
- MongoDB for persistent user sessions
- GCP auto-scaling + multi-language voice fallback to text

National Taiwan University — Insight AI Center | AI Engineer & Researcher

Jul 2024 – Feb 2025 | Taipei

- Won national AI championship (Taoyuan Business Innovation) — Multi-Agent + RAG system POC
- Delivered multiple industry POCs:
  - Automated market analysis reports
  - Industry knowledge retrieval system
  - Elderly anti-fraud AI (Smart City Hackathon — corporate support award + national finals)
- Led VLM + DPO reinforcement learning study group
- Published 2 papers at UC Berkeley AgentX

Tech: LangChain, LangGraph, Multi-Agent, RAG, GPT-4, Gemini, XGBoost, Streamlit, React, FastAPI, MongoDB, Qdrant, GCP

FunShine Group (Marketech International) | Medical AI Engineer

Jan 2023 – Jul 2024 | Taipei

- Built mammogram tumor detection (Vision Transformer, clinical-grade accuracy)
- Developed liver image segmentation for surgical planning
- Collaborated with Mackay Memorial Hospital on ML project
- Enhanced models through RLHF, DPO, GRPO reinforcement techniques
- End-to-end experience: R&D → clinical validation in regulated healthcare

Tech: PyTorch, Vision Transformer (ViT), Medical Imaging, CNN, RLHF, DPO, GRPO, Python

Earlier Roles

- ASE Group — Computer Vision Engineer (Jul 2022 – Jan 2023)
  - Deployed YOLO/CNN defect detection on semiconductor production lines
  - Automated visual inspection; reduced manual time significantly
- Shuttle Inc. — AI Product Manager (May 2021 – Jun 2022)
  - First AI role; managed AI product lifecycle for enterprise clients
  - Taiwan AI Academy Executive Program (Class 9)
- NSRRC — Production Line Manager (Prior to 2020)

  - Managed synchrotron radiation facility; built systematic analytical thinking

SIDE PROJECT

Casper AI Product Image Designer | Freelance | Solo Full-Stack Developer

Aug – Dec 2025 | Taipei

- Built AI-powered e-commerce product image platform as freelance project
- Full business cycle: client negotiation → development → delivery → payment closure
- 5 AI modules:
  - Text-to-Image (10 aspect ratios × 3 resolutions × batch 1-4)
  - Image-to-Prompt | Reference Generation | Prompt Enhancement | AI Image Search
- Production pipeline: AI background removal → generation → compositing → batch export
- Scale: 15,000+ lines frontend, 5,000+ lines backend, 50+ components, 15+ API endpoints

Tech: React 18, Vite, Tailwind CSS, Framer Motion, FastAPI, Gemini API, Firecrawl, Remove.bg, SQLAlchemy, Docker

▸ **System Case Study**

1. What system did you build?

AI e-commerce product image platform with 5 Gemini multimodal modules, real-time canvas editor, gamified chatbot, and batch production.

2. How does it work?

- Upload product photo → AI removes background (Remove.bg)
- Describe desired scene via text or reference images
- Gemini generates background → Canvas API composites with shadow/lighting
- Batch export multiple variants
- AI Design Agent assists via natural language throughout

3. What problem did you solve?

E-commerce sellers spend hours per product image. Reduces production to minutes; non-designers can produce professional-quality images.

4. Key mechanisms?

- NanoBananaService (664 lines): multi-model switching + auto format detection
- React-RND drag-and-drop canvas with pixel-level ruler
- FloatingChatBot (1,063 lines): smart intent recognition (generate/search/analyze)
- Framer Motion animations + gamified UI

5. How do you ensure stability?

- Multi-model fallback: Gemini 2.5 Flash → Gemini 3 Pro
- Error retry logic in AI service layer
- Docker Compose containerized (Nginx + Uvicorn)
- Complete delivery docs (README, deploy guide, API docs, quick ref card)

TEACHING & ENTERPRISE SPEAKING

7+ institutions | Hundreds of students | Bridging technical AI and business strategy for executives.

- Deloitte: Enterprise AI Applications & No-code Agent workshops
- Hsinchu Science Park / Qpoint: LLM Applications & Prompt Engineering for HSIP managers
- National Tsing Hua University: Multi-Agent Systems & AI Development
- National Taipei University of Business: LLM Application Tools

- GenAI Stars Program: RAG Evaluation Workshop
- Vibe Coding Courses: Cursor, Codex, Claude Code

AWARDS & VALIDATION TRACK RECORD

CHAMPION	Taoyuan Business Innovation & Entrepreneurship — Multi-Agent + RAG System (2024)
AWARD	Taipei Smart City Hackathon — AI Multi-Agent Anti-Fraud for Elderly (Corp. Support + National Finals, 2024)
FINALIST	Tourism Bureau National Hackathon — Multi-Agent Smart Tourism + VIVE Eagle (2025)
FINALIST	Prudential Insurance Hackathon — XGBoost + LLM Prediction System (2024)
PUBLISHED	UC Berkeley AgentX — 2 Papers on Edge AI & Medical Multi-Agent Systems (2025)

INDUSTRY KNOWLEDGE

LegalTech: Contract review, consultation, compliance

E-commerce: AI product imaging, design automation

Semiconductor: Automated inspection, production line AI

Tourism: Smart tourism, AR, sustainability

ESG & Sustainability: ESG reporting, compliance, deep research

Healthcare: Medical imaging AI, VLM fine-tuning

Insurance & Finance: Risk prediction, data analytics

Commerce & Retail: Smart commerce, automated reporting

EDUCATION & CERTIFICATION

- Taiwan AI Academy — Executive Program (Class 9): ML, DL, AI business strategy with 300+ executives (2021)
- UC Berkeley AgentX Program: Online research & competition, 2 published papers (2025)
- Languages: Mandarin (Native), English (TOEIC 720)