Yaohui (Tomie) Zhang

AI Application Developer | LangChain & RAG Specialist

📍 Auckland, New Zealand (North Shore)
📞 +64 20 477 4618
📧 tomieweb@gmail.com
💼 LinkedIn: linkedin.com/in/tomiezhang
💻 GitHub: github.com/neilzhangpro
🏷️ **Visa Status:** Student Visa with full work rights (25 hours/week)
✅ **Availability:** Immediately available for part-time/contract work

---

PROFESSIONAL SUMMARY

AI Application Developer specializing in **LangChain, RAG systems, and LLM integrations**. Combining 16 years of software engineering experience with recent deep-dive into AI application development.

What I bring:

- Production experience building AI-powered applications (chatbots, RAG systems, multi-agent workflows)
- Strong understanding of LLM fundamentals (prompting, context management, token optimization)
- Full-stack capabilities: can build complete AI applications from frontend to backend
- Published author and educator in AI development (2,000+ students, published book)
- Active in Auckland's tech community (Digital Seniors & Code Club volunteer)

Ideal engagement: AI integration projects, chatbot development, RAG system implementation (15-25 hours/week)

TECHNICAL SKILLS

AI & LLM Engineering (Production-Ready)

- **LLM Frameworks:** LangChain, LangGraph, LangSmith (observability)
- **LLM APIs:** OpenAI (GPT-4, GPT-3.5), Anthropic Claude, Azure OpenAI

- **RAG Systems:** Vector databases (ChromaDB, Pinecone), embeddings, semantic search
- **AI Patterns:** Prompt engineering, few-shot learning, chain-of-thought, ReAct agents
- **Multi-Agent Systems:** CrewAI, AutoGen (familiar with agent orchestration patterns)

Full-Stack Development

- **Frontend:** React, Next.js 14, TypeScript (for building AI chat interfaces)
- **Backend:** Python (FastAPI, Uvicorn), Node.js (Express)
- **APIs:** RESTful, WebSocket (for streaming LLM responses)
- **Databases:** PostgreSQL, MongoDB, Vector DBs (ChromaDB, Pinecone)

DevOps & Tools

- **Containerization:** Docker, Docker Compose
- **CI/CD:** GitHub Actions
- **Cloud:** AWS (EC2, S3), basic Azure experience
- **Monitoring:** LangSmith, custom logging solutions

RECENT WORK EXPERIENCE

AI Technical Consultant (Contract)

IronMind.ai | Remote | Oct 2024 – Nov 2025

Mental healthcare platform using AI to analyze therapy sessions and provide clinical insights

Tech Stack: LangChain, OpenAI GPT-4, ChromaDB, FastAPI, Next.js

Key Contributions:

1. RAG System for Clinical Insights

- Designed and implemented a RAG pipeline to analyze therapy session transcripts and generate evidence-based recommendations for psychologists
- **Technical Details:**
  - Used LangChain's `RetrievalQA` chain with custom prompt templates

    - Embedded 10,000+ clinical psychology research papers using OpenAI `text-embedding-ada-002`
    - Stored embeddings in ChromaDB with metadata filtering (by therapy type, patient demographics)
    - Achieved 85% relevance score in clinical accuracy testing (validated by licensed psychologists)

2. Conversational Memory Management

- Built a long-term conversation memory system to maintain context across multiple therapy sessions
- **Technical Details:**
    - Implemented custom `ConversationBufferWindowMemory` with sliding window (last 10 messages)
    - Used `ConversationSummaryMemory` for long-term context compression (reduced token usage by 60%)
    - Stored conversation summaries in PostgreSQL with vector similarity search for related past sessions

3. Multi-Agent Workflow

- Created a multi-agent system for automated session analysis:
    - **Transcription Agent:** Whisper API for audio-to-text
    - **Analysis Agent:** GPT-4 for sentiment analysis and key theme extraction
    - **Recommendation Agent:** RAG-based clinical suggestions
- Orchestrated agents using LangGraph with human-in-the-loop approval gates

4. Production Deployment

- Set up FastAPI backend with async endpoints for LLM streaming responses
- Implemented rate limiting and token usage monitoring (LangSmith integration)
- Deployed on AWS EC2 with Docker, achieving 99.5% uptime

Impact:

- Reduced psychologists' post-session documentation time by 40%
- Secured HK$500K incubation funding (Cyberport Hong Kong)

AI Educator & Content Creator

Self-Employed | Remote | Jan 2023 – Oct 2024

Online Course: "AI Agents from 0 to 1"

- Created comprehensive AI development course on imooc.com (Chinese MOOC platform)
- **Enrollment:** 2,000+ students, 4.8/5.0 rating
- **Topics Covered:**
  - LangChain fundamentals (chains, agents, memory)
  - Building production RAG systems
  - Multi-agent orchestration with CrewAI
  - Deployment best practices (Docker, FastAPI, streaming)

Published Book: "AI Agent Development Guide"

- Authored 300-page technical book (China Water & Power Press, 2025)
- **Content:** Step-by-step guide to building LLM applications with code examples
- **Reception:** 5,000+ copies sold in first 3 months

Open Source Projects:

- Published 8 AI demo projects on GitHub (total 500+ stars)
- Examples include:
  - RAG chatbot with document upload (LangChain + Pinecone)
  - Multi-agent research assistant (CrewAI)
  - Streaming chat interface (Next.js + FastAPI + WebSocket)

Technical Consultant (Part-time)

Multiple Clients | Remote | 2023 – 2024

Ironmind.ai - AI Technical Advisor

- Advised on LLM integration strategy for mental health platform
- Conducted technical due diligence for investor presentations

Haikou Kuidou Culture - AI Consultant

- Designed AI-powered content generation system for cultural tourism marketing
- Implemented GPT-4 based copywriting assistant (reduced content creation time by 50%)

Xi'an Tourism Group - Senior Advisor

- Consulted on smart tourism platform architecture
- Evaluated AI vendors for chatbot and recommendation systems

Software Engineering Background

Tencent | Shenzhen, China | Sep 2007 – Dec 2023

16 years of software engineering experience (frontend, full-stack, technical leadership)

Relevant to AI Work:

- Built high-scale web applications serving 300M+ users (experience with performance optimization)
- Led technical teams of 50+ engineers (project management, architecture design)
- Managed ToB products with $10M+ annual revenue (understanding of business requirements)

Key Takeaway: Strong software engineering foundation enables me to build production-ready AI applications, not just prototypes.

EDUCATION

Master of Software Engineering
Yoobee College of Creative Innovation, Auckland, NZ | Feb 2025 – Nov 2026 (In Progress)
Focus: AI Application Development, Cloud Architecture

Master of Software Engineering
Beihang University, Beijing, China | 2017 – 2020

AI PROJECT PORTFOLIO

GitHub: github.com/neilzhangpro

Featured Projects:

1. **RAG-Powered Documentation Chatbot**
   - Tech: LangChain, OpenAI, Pinecone, Next.js
   - Features: Document upload, semantic search, source citation
   - [Live Demo] | [Code]
2. **Multi-Agent Research Assistant**

    - Tech: CrewAI, GPT-4, Serper API (Google Search)
    - Features: Automated research, fact-checking, report generation
    - [Code]
3. **Streaming Chat Interface**
    - Tech: FastAPI, WebSocket, React
    - Features: Real-time LLM response streaming, token usage tracking
    - [Live Demo] | [Code]

Available upon request:

- Detailed technical documentation
- Code walkthroughs
- Client references

COMMUNITY INVOLVEMENT

Digital Literacy Volunteer
Digital Seniors, Auckland | Dec 2024 – Present

- Teaching elderly residents technology skills
- Demonstrates ability to explain complex concepts clearly

Code Club Volunteer
Code Club Aotearoa | Dec 2024 – Present

- Teaching children programming basics
- Every Tuesday, 2-hour sessions

WORK PREFERENCES

- **Hours:** 15-25 hours/week (flexible schedule)
- **Work Mode:** Remote preferred, open to hybrid (North Shore)
- **Rate:** $60-90/hour (depending on project complexity)
- **Start Date:** Immediately available
- **Ideal Projects:**
    - RAG system implementation
    - Chatbot development (customer service, internal tools)
    - LLM integration into existing applications
    - AI workflow automation

    - Technical consulting on AI strategy

WHY HIRE ME FOR AI PROJECTS?

✅ **Production Experience:** I've shipped real AI applications, not just tutorials
✅ **Full-Stack Capability:** I can build the entire AI application (frontend + backend + deployment)
✅ **Cost-Conscious:** I understand token optimization and cost management for LLM applications
✅ **Clear Communicator:** Published author and educator, can explain technical decisions to non-technical stakeholders
✅ **Local Presence:** Based in Auckland, available for in-person meetings if needed

---

CERTIFICATIONS & LEARNING

- **Currently Pursuing:** AWS Certified Solutions Architect (planned completion: Q1 2025)
- **Completed:** DeepLearning.AI LangChain courses (Coursera)
- **Active Learning:** Following latest AI research (daily reader of arXiv, Hugging Face papers)

REFERENCES

Available upon request (including reference from IronMind.ai project lead and course students)