Jiechao Cai

BSc in Computer Science and Technology | Incoming MSc in Computer Science at HKU | AI Agent / LLM Application Development

2651159710@qq.com (+86) 136-0270-5148

www.caijiechao.com github.com/computersniper jiechao-cai

Target Positions: AI Agent Engineer, LLM Application Engineer, LLM Platform Developer

Education

Beijing Normal-Hong Kong Baptist University (BNBU) & University of Malaya (Exchange) Sep. 2022 – Jun. 2026

BSc in Computer Science and Technology English-taught program; Top 15% in major; Second-Class Scholarship

- Core Courses: Data Structures and Algorithms, Operating Systems, Data Communications and Networks, Machine Learning, Natural Language Processing, Deep Learning and Neural Networks.

The University of Hong Kong Sep. 2026 – Aug. 2027 Expected

MSc in Computer Science

- Planned focus: Large Language Models and Natural Language Processing, AI/Agent Applications, Reinforcement Learning and Alignment, and Model Fine-tuning.

Technical Skills & AI Strengths

- **Advanced AI-Assisted Development**: Heavy and experienced user of Cursor, Claude Code, and Trae, with daily high-volume LLM-assisted development experience across multiple paid AI coding platforms. Skilled in prompt design, prompt debugging, and using LLMs to generate production-level code efficiently.
- **Agent and LLM Engineering**: Strong understanding of LLM capabilities and limitations. Experienced in Context Engineering, RAG knowledge base construction, MCP-based Agent integration, and Skill development for practical Agent deployment.
- **Core Technical Stack**: Proficient in Python; familiar with Java, TypeScript, and Kotlin. Experienced with FastAPI, SQLAlchemy, Docker, asyncio, PostgreSQL, Redis, and full-stack development. Solid foundation in data structures, algorithms, computer networks, and backend engineering.

Internship & Project Experience

Rhino International Education | AI Platform R&D Intern Jan. 2026 – Mar. 2026

- Worked at an education content service provider that supplies high-quality Olympiad-level mathematical problem data to AI companies such as Tencent, Alibaba, and Xiaohongshu.
- Served as a full-stack developer responsible for frontend and backend architecture, core feature development, and testing. Built data cleaning and intelligent annotation pipelines, integrated LLMs for problem quality evaluation, and developed a LiteLLM gateway for performance optimization, usage tracking, and automated daily reporting.

MathTasks: Intelligent Math Problem Production Platform Jan. 2026 – Feb. 2026

Python, FastAPI, Next.js, PostgreSQL, Vector Search, Playwright, Agent Development, LLMs & Multimodal LLMs

Link: stem-align.com

- **Core Architecture and Workflow**: Designed a finite state machine (FSM) and RBAC-based workflow covering 5 user roles, resolving concurrent task transition conflicts. Built a math problem database model with 30+ fields and a 5-dimensional automatic scoring system.
- **Asynchronous Queue and Real-Time Updates**: Built asynchronous task queues for difficulty evaluation and vector-based duplication detection, supporting efficient deduplication for large-scale problem banks. Integrated SSE streaming with Redis-based state caching to provide real-time progress updates and recovery after frontend disconnection.
- **AI Quality Evaluation**: Designed an LLM-based adversarial validation mechanism using multi-sampling and threshold checking, converting probabilistic LLM outputs into more deterministic difficulty evaluation signals.
- **Automated Testing**: Implemented multi-role end-to-end testing with Playwright. Integrated multimodal LLMs for CAPTCHA recognition and designed Tesseract/manual-intervention fallback strategies to improve the robustness of the main workflow.

Enterprise-Level Unified LLM Gateway and API Management Platform Feb. 2026 – Apr. 2026

Python, LiteLLM, Docker, PostgreSQL, Redis, Asyncio, Langfuse

- **Gateway Routing and High Availability**: Unified access to 20+ models from 8+ providers, including OpenAI, Anthropic, and Doubao. Configured multi-key load balancing and automatic failover to handle upstream rate limits and

network instability.

- **Performance Testing and Optimization**: Developed concurrent load testing scripts to evaluate and optimize first-token latency and throughput under 50+ concurrent requests. Tuned long-running reasoning connection settings to support thousands of complex Agent tasks per day.
- **Cost Governance and Observability**: Integrated Langfuse for token-level tracing. Developed scheduled automation scripts to parse and aggregate gateway logs, generating fine-grained cost reports by team and project and sending automated daily reports via Feishu Webhook.
- **Scale and Traffic**: Processed over 100 million tokens in total and stably supported thousands of complex reasoning calls per day.

Beijing Normal-Hong Kong Baptist University | Teaching Assistant Sep. 2025 – Dec. 2025

- Served as a teaching assistant for the course *Data Structures and Algorithms*, responsible for tutorial sessions, assignment grading, lab guidance, and helping students understand core algorithmic concepts.
- Led the development of a Multi-Agent AI teaching assistant system. Extracted structured knowledge from course materials to build a RAG-based knowledge base, enabling automated answers for over 50% of common student questions.