CHEN Yixuan (Ethan)

hi@yixuan.app | LinkedIn

Applied AI undergraduate at HKU (CGPA 4.19/4.30) with experience tuning models, deploying LLM agents, RAG systems, and scalable backend services. I am interested in applying AI engineering to large-scale, production-level, real-world applications. Available for summer internships from late May 2027.

EDUCATION

The University of Hong Kong **Sep 2024 – Jun 2028 (expected)**
Bachelor of Arts and Sciences in Applied Artificial Intelligence | CGPA: 4.19 / 4.30

- **Entrance Scholarship for Outstanding Academic Talents (HK$182,000/year)**
- Coursework: Programming I & II (A+), Data Structures and Algorithms (A), Intro to Data Science (A+), Intro to AI (A+), Database Systems (A+)

EXPERIENCE

WYNI Technology Limited / *AI Engineer Intern* **Nov 2025 – Jun 2026**

- Shipped a Cantonese–English meeting-minutes pipeline used in production by **a top-10 Hong Kong insurer**: speaker diarization, energy-based segmentation, a Cantonese-tuned ASR model with domain prompting, and LLM structured output into minutes templates; **cut transcription error rate from 27.9% to 9.3%** vs. WhisperX (large-v3) on 266 meeting clips.
- Built a LangChain agent that turns raw Outlook claim emails into Excel workbooks ready for review (email and attachment parsing, evidence extraction, schema-based analysis), **cutting per-case preparation time from 23 to 5 minutes**.
- Re-architected company's agent tool integration around a progressive-disclosure MCP gateway (3 fixed meta-tools, schemas loaded on demand); across 83 tools, **cut per-task input tokens by 92%+ and cost by 70%+**, with zero tool-call errors on a 60-case eval set (40 deliberately confusable).

Visual AI Lab, the University of Hong Kong / *Research Assistant* **Jul 2026 – Present**

- Building an X-ray contraband screening pipeline with **Hong Kong International Airport** (detector → LLM report → claim verification), owning the stage that turns detections into Chinese risk reports with per-claim confidence for screeners.
- Created a **200-case hallucination benchmark** (20 injected error types) to select the verifier: TypeSafe Jev, a text-only model scoring each claim against upstream evidence, reached 0.988 AUROC (vs. 0.92 open-source).

PROJECTS

FirmGround Agentic RAG – Domain-agnostic Agentic RAG Platform *PostgreSQL, OpenSearch, Redis, React*
Piloting as the history knowledge base (textbooks, curriculum references) of a public junior secondary school in Tianjin.

- Designed a ReAct agent that rewrites queries, decomposes multi-hop questions, and decides when evidence suffices, using hybrid retrieval (BM25 + dense, RRF, Qwen3 reranker) as tools; on FinanceBench 150 questions open set, **reached 81.33% accuracy vs. 21.33%** for naive RAG with the same model and embeddings.
- Engineered native-async document ingestion and agent execution with FastAPI, Taskiq, PostgreSQL, and Redis Streams (resumable SSE, event coalescing); **ingests 372 pages/min** across concurrent uploads within 650 MiB peak memory.

FirmGround Ledger – AI Bookkeeping & Reconciliation *FastAPI, LangChain, Redis, React*
Reconciliation workbench for accountants: a VLM turns receipts into ledger entries, which an algorithm matches against bank statements and ERP exports, with agents reserved for ambiguous cases.

- Designed a gated agentic workflow in which a deterministic algorithm runs ingestion, exact/tolerance/fuzzy/subset matching, and break classification; human-approved format mappings become versioned templates, so **known formats are ingested with zero LLM calls**.
- Built a multi-agent layer of three specialized ReAct agents (format mapping, match review, root-cause investigation), each sandboxed in an isolated context; human-in-the-loop gates keep their outputs as suggestions until an accountant confirms, and confirmed decisions replay deterministically for a full audit trail.
- On 10 multi-page statements, **cut processing time by 60%+** vs. manual Excel bookkeeping, including time for human review and correction.

AWARDS AND ACHIEVEMENTS

- **Dean's Honours List (2025, 2026)** for academic excellence, awarded by Faculty of Science, HKU
- **Ninestar Scholarship** (2024) & **Winnie S M Tang Scholarship** in Applied AI (2026)
- **Student Adviser** for Applied AI Program; presenter at multiple student-led GenAI seminars

TECHNICAL SKILLS

Languages: Python, C++, R, SQL, JavaScript
AI/LLM: agentic workflows, multi-agent systems, context engineering, agent harnesses, RAG, LLM evaluation, human-in-the-loop design, LoRA fine-tuning, prompt engineering
Frameworks & Tools: PyTorch, LangChain, LangGraph, MCP, Agent Skills, OpenAI SDK, Claude Code
Backend & Infra: FastAPI, PostgreSQL, OpenSearch, Redis, Taskiq, Docker, Git

LANGUAGES

Putonghua (native) | English (fluent, IELTS 8) | Cantonese (fluent)