Haodong Zhang

Email: 1010761880@qq.com | **Phone:** (+86) 131-4879-6380 | **GitHub:** holeeast979.github.io

Education

The Chinese University of Hong Kong *MSc in Artificial Intelligence* Incoming Fall 2026

Jinan University (JNU) *B.Eng. in Software Engineering* Sept. 2022 - June 2026

- **GPA:** 3.85/5.0 (Top 5%) **Awards:** National Scholarship (Top 1%), Blue Bridge Cup (Provincial 2nd Prize)
- **Core Courses:** Reinforcement Learning (97), Machine Learning (96), AI Principles (92), Algorithm Analysis (90)

Technical Skills

- **Languages:** Python (Expert), C/C++, Java, SQL **AI Frameworks:** PyTorch, LangChain, LangGraph, Hugging Face
- **LLMOps & Tools:** Docker, Langfuse, Prometheus, Grafana, RocketMQ, Git, Linux
- **Core Competencies:** LLM App Dev (RAG/Agents), LMM Profiling & Architecture Analysis

Internship Experience

Shifangronghai Technology Co., Ltd. AI Platform Developer Intern (LLMOps) Sept. 2025 - Dec. 2025

- **Evaluation Pipeline:** Containerized high-concurrency evaluation suite via Docker/RocketMQ. Developed Mock Server to simulate massive requests, establishing TPS & Latency baselines for production readiness.
- **Observability:** Integrated Langfuse into LangGraph, enabling end-to-end tracing from queries to tokens. Solved ”black box” debugging issues in multi-agent systems, reducing error attribution time by 50%+.
- **Monitoring:** Built Prometheus/Grafana dashboards tracking ”Queue Lag” and ”Consumption Latency” for millisecond-level anomaly alerting.
- **Stack:** Python, LangGraph, Docker, RocketMQ, Prometheus, Grafana, Langfuse

Project Experience

Research on Edge-side LMM Inference Acceleration Independent Lead Oct. 2024 - Present

- **Bottleneck Profiling:** Developed non-intrusive profiler via PyTorch Hooks. Identified ViT Encoder consumes 92-94% of latency, debunking video decoding bottleneck assumptions (only 18%).
- **Architecture Analysis:** Validated Q-Former superiority for video tasks, achieving 9x speedup (923ms $\rightarrow$ 105ms) vs. MLP architectures via token compression.
- **Sampling Strategy:** Revealed high frame rate ”diminishing returns” (77% latency hike vs. negligible accuracy gain), establishing dynamic key-frame extraction strategy.
- **Optimization (Ongoing):** Researching FlashAttention edge adaptation and Token Pruning for 40%+ speedup.
- **Stack:** PyTorch, Hooks, Profiling (Nsight Systems), Transformers, ViT

High-Quality RAG Benchmark Construction Core Developer June 2025 - Aug. 2025

- **Data Engineering:** Built automated augmentation pipeline using DeepSeek API. Expanded ”Golden Passages” by 214% (13k $\rightarrow$ 42k) for NQ dataset, improving long-tail coverage.
- **Retrieval Optimization:** Implemented Re-ranking strategy, improving Recall@10 by 3-5% in benchmarks.
- **Stack:** Python, DeepSeek API, Milvus, Spacy, Data Processing

Honors & Awards

- **National Scholarship (Top 1%):** Highest undergraduate honor | Jinan University, 2025
- **Winner Prize, APMCM:** Core member (LLM Track) | 2025
- **Provincial 2nd Prize, Blue Bridge Cup:** Python Track A | 2024
- **Outstanding Student Scholarship:** Awarded consecutively | 2022-2024