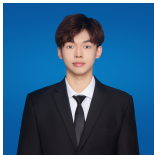


# Haodong Zhang

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
<ul style="list-style-type: none"><li><b>GPA:</b> 3.85/5.0 (Top 5%)    <b>Awards:</b> National Scholarship (Top 1%), Blue Bridge Cup (Provincial 2nd Prize)</li><li><b>Core Courses:</b> Reinforcement Learning (97), Machine Learning (96), AI Principles (92), Algorithm Analysis (90)</li></ul>		

## 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
<ul style="list-style-type: none"><li><b>Evaluation Pipeline:</b> Containerized high-concurrency evaluation suite via Docker/RocketMQ. Developed Mock Server to simulate massive requests, establishing TPS &amp; Latency baselines for production readiness.</li><li><b>Observability:</b> 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%+.</li><li><b>Monitoring:</b> Built Prometheus/Grafana dashboards tracking "Queue Lag" and "Consumption Latency" for millisecond-level anomaly alerting.</li><li><b>Stack:</b> Python, LangGraph, Docker, RocketMQ, Prometheus, Grafana, Langfuse</li></ul>		

## PROJECT EXPERIENCE

Research on Edge-side LMM Inference Acceleration	Independent Lead	Oct. 2024 - Present
<ul style="list-style-type: none"><li><b>Bottleneck Profiling:</b> Developed non-intrusive profiler via PyTorch Hooks. Identified ViT Encoder consumes 92-94% of latency, debunking video decoding bottleneck assumptions (only 18%).</li><li><b>Architecture Analysis:</b> Validated Q-Former superiority for video tasks, achieving 9x speedup (923ms → 105ms) vs. MLP architectures via token compression.</li><li><b>Sampling Strategy:</b> Revealed high frame rate "diminishing returns" (77% latency hike vs. negligible accuracy gain), establishing dynamic key-frame extraction strategy.</li><li><b>Optimization (Ongoing):</b> Researching FlashAttention edge adaptation and Token Pruning for 40%+ speedup.</li><li><b>Stack:</b> PyTorch, Hooks, Profiling (Nsight Systems), Transformers, ViT</li></ul>		
High-Quality RAG Benchmark Construction	Core Developer	June 2025 - Aug. 2025
<ul style="list-style-type: none"><li><b>Data Engineering:</b> Built automated augmentation pipeline using DeepSeek API. Expanded "Golden Passages" by 214% (13k → 42k) for NQ dataset, improving long-tail coverage.</li><li><b>Retrieval Optimization:</b> Implemented Re-ranking strategy, improving Recall@10 by 3-5% in benchmarks.</li><li><b>Stack:</b> Python, DeepSeek API, Milvus, Spacy, Data Processing</li></ul>		

## 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