Yeqi Huang

Email: yeqi.huang@ed.ac.uk

Personal Site: yeqi-huang.com

EDUCATION

- **University of Edinburgh** — Edinburgh, UK
  PhD in Computer Science — *August 2023 - Present*
  Research Field: *AI-Systems, Distributed ML, ML-oriented Architecture, Serverless*

- **University of Science and Technology of China** — Hefei, China
  Bachelor of Computer Science; The school of Gifted Young; — *July 2017 - June 2021*
  Courses: *Operating Systems, Artificial Intelligence, Principles of Compiler, Computer Architecture, High Performance Architecture*

PUBLICATIONS

1 Yao Fu, Leyang Xue, **Yeqi Huang**, et al. "ServerlessLLM: Locality-Enhanced Serverless Inference for Large Language Models." In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, 2024.

2 Yao Fu, Yinsicheng Jiang, **Yeqi Huang**, et al. "MoE-CAP: Cost-Accuracy-Performance Benchmarking for Mixture-of-Experts Systems." *NeurIPS*, 2025.

3 Congjie He, **Yeqi Huang**, Pei Mu, et al. "WaferLLM: Large Language Model Inference at Wafer Scale." *Under Review*, OSDI 2025.

4 Xiao-Long Chen, Lin-Feng Wang, **Yeqi Huang**, et al. "Symplectic structure-preserving particle-in-cell whole-volume simulation of tokamak plasmas." In *SC21: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021.

5 Yinsicheng Jiang, **Yeqi Huang**, Liang Cheng, et al. "RAGBoost: Efficient Retrieval-Augmented Generation with Accuracy-Preserving Context Reuse." *arXiv preprint*, arXiv:2511.03475, 2025.

PROJECTS

- **QED Simulation on GPU**:
  - Implemented high-performance Monte-Carlo simulation achieving 1500x to 5000x speedup.
  - Received recognition from Nobel Laureate Frank Wilczek at APS conference.
  - Demonstrated expertise in CUDA programming and GPU micro-architecture optimization.

- **AutoReader Project**:
  - Built a vector database for the latest ArXiv papers for daily semantic search and subscription.(Full-stack AI application)
  - Achieving **60× faster indexing than NVIDIA cuVS** library and **13× faster retrieval than Milvus**.
  - Extended system to support bioRxiv and PubMed papers with specialized biology-focused features.
  - Biology patched version is demostrating in Dartmouth university, will submit to Nature this year.

- **MICA (submitted to OSDI 2026)**:
  - Developed the first end-to-end compiler stack for mesh accelerators (Cerebras, Tenstorrent) with a novel MeshIR.
  - Implemented a search framework reducing schedule search from weeks to hours for millions of cores.
  - Achieved 3.6× performance improvements over state-of-the-art compilers and 1.4× over expert hand-tuned codes.
  - Enabled applications to run 100–200× faster than on GPUs by efficiently exploring temporal-spatial schedules.

- **SwarmX (submitted to OSDI 2026)**:
  - Presented the first scheduler agent framework that formulates cluster scheduling as an agentic intelligence problem.
  - Deployed in production on tens of thousands of servers, improving P99 latency by up to 50% and doubling throughput.
  - Designed modular agents with neural predictors and memory to optimize complex workflow performance.
  - Maintained compatibility with Ray and ComfyUI, enabling existing workflows to run without modification.

- **BioVLM Research Project**:
  - Collaborated with Dartmouth and Harvard Medical School on biological literature interpretation.
  - Developed advanced Vision ability by adapting GRPO RL Training, CoT training, ReFT, and long context with RAG.
  - Achieved 23% improvement over GPT-4o in double-blind human evaluations.
  - Achieved 10% improvement over GPT-5 in LabBench benchmark.

- **AI4Math - Sketchpad Project**:

- Received grant from AI for Math Fund to develop Sketchpad system for formal mathematics (announcement)
  - Building system that automatically converts mathematical proofs into structured data using graph-based representation.
  - Enabling decomposition of proofs into individual statements for more precise auto-formalization.
- **More Projects (https://yeqi-huang.com/)**:
  - I have plenty of AI related projects on my personal page.

Talks

1. **Yeqi Huang**. "InfiniTensor: A Tensor-Friendly, Efficient Parallel Programming Library for Accelerator-Centric Clusters." *UKSys 2024*.

2. **Yeqi Huang**. "Why we need a new clustering benchmark in AI retrieval?" *International Workshop on Efficient Generative AI*, 2024.

Teaching Experience

- **Machine Learning System TA**:
  - Teaching assistant for CUDA programming course, covering GPU architecture, CUDA, CuPy, and Triton.
  - Designed hands-on exercises and projects to help students master parallel programming concepts.

Awards

International Awards:

- International Supercomputing Conference Student Cluster Competition Champion - 2021
- Asian Supercomputer Conference Student Cluster Competition First Prize - 2021

National Awards:

- Best Chinese Supercomputing Application of the Year - 2022
- Huawei Pioneer Developer (4 in China) - 2021
- National Compiler Designing Competition Second Prize - 2021

Skills

- **AI-Related**:
  - In-depth understanding of LLMs, with experience porting high-performance inference and training frameworks to various hardware platforms, including Apple Silicon, GPUs, and Cerebras.
  - Strong understanding of vector retrieval, having developed and tested graph and vector databases on multiple platforms.
  - In-depth knowledge of Multi-Agent applications and developed efficient development components for such applications.

- **Computer Science Related**:
  - Highly skilled in **C++, Python, Rust**; familiar with Go, JavaScript, and Latex
  - Proficient with **CUDA, Intel ONEAPI, OpenMP** and related parallel and distributed programming
  - Well-versed in **LLVM**, frequently participating in LLVM Forum online discussions
  - Extensive experience working with Linux, including usage of eBPF
  - Rich experience in distributed systems and distributed machine learning frameworks
  - Strong engineering development experience, mastering various compiler-related tools and static analysis tools

- **Physics & Math**:
  - Highly skilled in Computational Fluid Dynamics and Molecular Dynamics
  - Strong knowledge in numerical methods and linear algebra
  - Proficient in Quantum Mechanics and Quantum Electrodynamics, with some knowledge of quantum computing

Special Experience

- **Open Source Enthusiast**: Contributing to notable projects like GNOME, LAMMPS, and LLVM, I've used GitHub since 2019 to showcase my development journey and ideas.

- **Running a Science-Themed Cafe**: Leveraging software development income, I established Quantum Coffee near my school —a collaborative space encouraging students to explore and discuss scientific topics across disciplines.

- **UNICEF Charity Projects**: I have donated 10% of all personal income to children's charities since 2019.