Marshall Yiding Wang

LLM researcher and engineer
wydwww@gmail.com | yidingwang.xyz | Hong Kong

EDUCATION

Hong Kong University of Science and Technology **April 2023**
Ph.D. in Computer Science Hong Kong
- Thesis: Towards Efficient Deep Learning Systems with Learning-Based Optimizations.

Shanghai Jiao Tong University **July 2017**
B.Eng. in Instrument Science (Electrical Engineering). Shanghai, China

WORK

Huawei AI Lab **Sep 2023 – Now**
Researcher Hong Kong
- Working at the NLP and Multimodal group. Top 15% performance in 2025.
- **MoE**: Led an MoE (mixture of experts) training project using an improved router to achieve better expert selection and sparse activation for accuracy and efficiency. Deployed a suite of data processing, training, and evaluation pipelines in production.
- **AI Tools**: Experienced with agentic coding in production. Delivered projects like codebase migration, ML training/eval dashboard, data quality validation dashboard, and RAG solution for finance documents.
- **Pre-Training**: Trained LLMs at scale with 1000+ GPUs. Built a scaling law benchmark that aligns with OpenAI and enables accurate cost–performance projections for future LLM training investments. Conducted LLM experiments with model arches, hyperparameters, datasets, and benchmarks.
- **Evaluation**: Deployed LLMs in development at scale for efficient evaluation with dozens of NLP and multimodal benchmarks to ensure performance across question answering, math, and visual understanding.
- **Framework Dev**: Worked with infra team to improve efficiency and maintain feature consistency of in-house training framework based on Nvidia's Megatron and MS's DeepSpeed on non-GPU computing platform.
- **Stakeholder Management**: Experienced with internal/external client management (engineers and managers) and cross-team development.
- **Tech Skills**: Python, PyTorch, HuggingFace, NLP, distributed training, vLLM, agentic coding.

Intel **Aug 2016 – Jan 2017**
Software Engineer Intern Shanghai, China
- Implementing Torch-like APIs, such as confusion matrix for Intel BigDL, a Spark-based DL framework.

PUBLICATIONS

Scaling Law for Language Models Training Considering Batch Size.

Xian Shuai, **Yiding Wang**, Yimeng Wu, Xin Jiang, Xiaozhe Ren**. arXiv 2412.01505**
- Training GPT-like models ranging from 130M to 2.6B parameters using up to 300 billion high-quality tokens and establishing a basic scaling law on model size and training data/compute.
- Analyzing how varying batch sizes and learning rates affect the convergence and generalization of LLMs.
- **Relevance:** Provides cost–performance optimization strategies for enterprise-scale LLM training.

Tabi: An Efficient Multi-Level Inference System for Large Language Models.

Yiding Wang, Kai Chen, Haisheng Tan, Kun Guo. **EuroSys 2023**

- Reduced inference latency of Transformer-based LLMs for text classification by 21–40% through confidence calibration, token pruning, and model ensembling.
- **Relevance:** Improves real-time NLP applications such as document classification.

Egeria: Efficient DNN Training with Knowledge-Guided Layer Freezing.

Yiding Wang, Decang Sun, Kai Chen, Fan Lai, Mosharaf Chowdhury. **EuroSys 2023**

- Accelerated training by 19–43% without sacrificing accuracy using knowledge distillation and transfer learning.
- **Relevance**: Speeds up LLM development, reducing training cost.

Enabling Edge-Cloud Video Analytics for Robotic Applications.

Yiding Wang, Weiyan Wang, Duowen Liu, Xin Jin, Junchen Jiang, Kai Chen. **INFOCOM 2021** and **TCC 2022**

- Designed an edge–cloud analytics system using super-resolution for bandwidth-constrained environments.
- **Relevance:** Demonstrates scalable AI system design, applicable to latency-sensitive data pipelines.

OTHER

- **Teaching Assistants:** Operating System (Spring 2020), Introduction to Computing with Java (Fall 2018)
- **Languages**: English (fluent), Cantonese (conversational), Mandarin (native)
- **Awards**: HKUST Ph.D. Studentship (2017-2023), ACM/IEEE/HKUST Conference Grants (5×)