Wanqi Yang

yangwq177@gmail.com
https://github.com/Superone77

EDUCATION

Technical University of Darmstadt, Germany Sep 2020 - Sep 2023

Computational Engineering Master

- GPA: 1.8/5.0 (Scale: 1.0 Best)
- Courses: Statistical Machine Learning, Robot Learning, Numerical & Statistical Methods, Visual Computing, Optimization, System & Parallel Programming, Programming Massively Parallel Processors, Programming in Automatic Control, Data Science, Data Structures & Algorithms, Information Management

Beijing Jiaotong University, China Sep 2015 - Jul 2019

Mechanical Engineering (Robotics Specialization) Bachelor

- GPA: 91.3/100 (top 2%)

RESEARCH INTERESTS

My research goal is to advance foundation-model systems from multiple perspectives: improving model capabilities, developing deeper understanding of model behavior, and making ML systems more efficient, reliable, and accessible. I like to take a whole-stack view, from representations and learning algorithms down to inference systems and low-level hardware backends.

PUBLICATIONS

Wanqi Yang, Yuexiao Ma, Alexander Conzelmann, Xiawu Zheng, Michael W. Mahoney, T. Konstantin Rusch, and Shiwei Liu. "AlphaQ: Calibration-Free Bit Allocation for Mixture-of-Experts Quantization." Under Review, 2026. [Paper] [Code]

RESEARCH EXPERIENCE

Empirical Study of Conditioning Structure for Efficient MoT-based Multimodal Model Mar 2026 - Present

Advisor: Dr. Yuexiao Ma and Dr. Xiawu Zheng

- Empirically studied BAGEL-7B-MoT by profiling how visual/text conditioning signals evolve across CFG branches, roles, layer depths, and denoising timesteps, revealing structure in conditioning lifetimes that motivates two inference-optimization directions: role-aware KV-cache management and condition-transport-aware mixed-precision quantization.

AlphaQ: Calibration-Free Bit Allocation for MoE Quantization Aug 2025 - Present

Advisor: Dr. Shiwei Liu and Dr. Yuexiao Ma | arXiv: 2606.04980 | Code: https://github.com/Superone77/AlphaQ

- Identified calibration-domain bias in data-driven MoE quantization, where bit allocation overfits to calibration data and hurts cross-domain generalization; proposed AlphaQ, a calibration-free method that estimates expert/layer importance from spectral structure and formulates global mixed-precision allocation as an ILP.

Numerical Interventions for Robust Low-Bit LLM Training May 2025 - Jan 2026

Advisor: Dr. Shiwei Liu and Dr. Zechun Liu | Code: https://github.com/Superone77/MultiBitsQ

- Studied low-bit LLM training as a numerical intervention on stability and generalization, developing an efficient FP4/NVFP4 framework with unified quantization modules, multi-bit noise injection, and stochastic bit-rotation (WiniQ/ParetoQ); achieved 3–4x FP4 forward acceleration and improved robustness of low-bit models such as MobileLLM.

Master Thesis: Quantization for Neural Networks based on Classification Difficulty Feb 2023 - Sep 2023

Advisor: Prof. Dr.-Ing. Grace Li Zhang and Jingcun Wang | Paper & Code: https://github.com/Superone77/ClassPathQ

- Studied classification difficulty as an empirical signal for adaptive neural network quantization, proposing a difficulty-aware precision allocation method to reduce computation and memory-access overhead while preserving accuracy; built a modular framework across VGG/ResNet and CIFAR-10/100, outperforming prior methods under the same compression ratio.

GPU Performance Modeling for Parallel Programs Apr 2021 - Jul 2021

Advisor: Dr. Stefan Guthe and Fabian Czappa | Code: https://github.com/Superone77/PMPP-Compositional-GPU

- Proposed and validated a performance modeling method for CPU/GPU parallel kernels, including matrix multiplication and vector operations, to analyze compute bottlenecks and guide optimization of heterogeneous programs.

WORK EXPERIENCE

Qualcomm Communications Technology Co., Ltd. | Machine Learning Engineer Dec 2024 - Present

- Studied accuracy-latency trade-offs of multimodal foundation model inference on Snapdragon mobile and PC platforms, developing full-stack adaptations across model structure, quantization, operators, speculative decoding, and pipelines.
- Validated the optimization pipeline on generative and vision-language models, including Stable Diffusion 3/3.5, Qwen-VL/Omni series, and Phi-4-Multimodal, through on-device benchmarks for reliable mobile/PC deployment.

DJI Technology Co., Ltd. | High Performance Computing Engineer - AI Inference Sep 2023 - Dec 2024

- Optimized heterogeneous inference for autonomous-driving perception models, including BEV and occupancy networks, on Qualcomm CPU/GPU/DSP platforms through preprocessing, custom operators, kernels, and hardware-aware pipeline design.
- Developed compiler passes and automated front-end processing to convert road-structure and 3D perception models into efficient executable representations across heterogeneous compute backends.

Technical University of Darmstadt, AI & Machine Learning Lab | Research Assistant Oct 2021 - Feb 2022

- Designed a cloud-based experimental environment for studying learning-agent behavior through multi-agent interactions, implemented 300+ evaluation experiments with deployment, logging, and server operations in multiplayer AI environments.

TECHNICAL SKILLS

- **Languages**: English (IELTS 7.0), German (C1, DSH 2), Mandarin Chinese (Native)
- **Programming**: Python, C/C++, CUDA, Triton, SQL, Java, Shell; parallel programming with MPI/OpenMP
- **Frameworks & Tools**: PyTorch, NumPy, Megatron-LM, Deepspeed, Docker, Git, Linux
- **Research & AI Systems**: Foundation models, multimodal generation, MoE quantization, low-bit training, efficient inference, speculative decoding, hardware-aware model optimization
- **Inference Optimization**: Model deployment and acceleration across NVIDIA GPUs, Qualcomm SoC, and Apple Silicon

SELECTED AWARDS

- Germany Graduation Thesis Scholarship
- Knorr Scholarship at Beijing Jiaotong University (Top 1%)
- First-class Academic Excellence Scholarship (Top 5%)
- First-class Social Work Scholarship (Top 5%)
- Second-class Academic Excellence Scholarship (Top 10%)
- Outstanding Undergraduate Thesis (Top 5%)
- National University Innovation Training Project Award (Top 5%)