Jeffrey Wong

+44 (0) 7493 693114 thw20@ic.ac.uk linkedin.com/in/jeffrey-wong-b7206b247/ github.com/JeffreyWong20

EDUCATION

Imperial College London **Oct. 2024 – June 2028**

PhD. Computer Engineering *London, UK*

- Advisor: Dr. Yiren (Aaron) Zhao, Prof. Wayne Luk
- Research: ML Optimization, LLMs Compression, Reasoning Knowledge Representation.

Imperial College London **Sep. 2020 – June 2024**

MEng. Electronic and Information Engineering, First-Class Honours *London, UK*

- Coursework: Advanced Computer Architecture, Advanced Deep Learning System, Machine Learning, Digital System

SKILLS

Programming Languages: Python, C++

PUBLICATIONS

A^3: an Analytical Low-Rank Approximation Framework for Attention **NeurIPS'25**

JTH.Wong, C.Zhang, X.Cao, P.Gimenes, GA.Constantinides, W.Luk, Y.Zhao *Under review*

- Proposed an analytical low-rank approximation framework customised for Attention, outperforming the previous SoTA's 7.87 by 3.18 perplexity under the same reduction budget
- Compressing the KV cache, reducing memory footprint, and minimizing FLOPs, all with 0 runtime overhead

QERA: an Analytical Framework for Quantization Error Reconstruction **ICLR'25**

C.Zhang, JTH.Wong, C.Xiao, GA.Constantinides, Y.Zhao

- Analyzed and developed an analytical closed-form solution to compute the optimal low-rank terms A_k and B_k that best reconstruct the quantization error of LLMs, outperforming the previous SoTA's by 6.05% of 2-bit RoBERTa on GLUE
- Evaluated post-training quantization performance of QERA on the 4-bit LLaMA-3.1-70B model, achieving a 2.97% improvement over ZeroQuant-V2

ARIES: Autonomous Reasoning with LLM on Interactive Thought Graph Environments **EMNLP'25**

P.Gimenes, Z.Cao, JTH.Wong, Y.Zhao *Under review*

- Examined a multi-agent architecture for autonomous planning and execution, leveraging graph environments to solve complex LLM queries

EXPERIENCE

Terra API (YC W21) **April 2023 – Oct. 2023**

Software Engineer *London, UK*

∗ Integrated 8 health device providers into the Terra API backend, allowing Terra to query data from new health devices
∗ Designed and developed Terra's mobile app to provide customers with an entry point to integrate with Terra
∗ Conducted market research and industry analysis for the development of Terra new product

PROJECTS

ASIC for Transformer | *Software-Hardware co-optimisation* **Oct. 2024**

∗ Analyzed data flow in LLaMA model computations to optimize on-chip data reuse and off-chip memory communication
∗ Implemented the FlashAttention v2 algorithm on ASIC using a custom instruction set architecture (ISA)

Benchmarking tool for alternative compute paradigm | *SNN, PIM, Optical Compute* **Jun. 2024**

∗ Developed transformation passes that transform a given model into a model with alternative compute paradigm: SNN, PIM, Optical compute

LEADERSHIP & AWARDS

Graduate Teaching Assistant **2022-2025**

∗ ELEC70109, ELEC50009, ELEC50006: Advanced Deep Learning System, Information Processing, Discrete Math

Hackathon **2022 - 2024**

∗ IC HACK'24: 3rd prize in trading, built a trading algorithm to compete in Optiver trading stimulation.
∗ IC HACK'25: 2nd prize in Application of Novel Research, built and trained a quantized neural network with RL to play Snack/Tetris game on a memory constraint system with Raspberry Pi RP2040

Academic Awards **2022**

∗ Awarded 2nd place in the Head of Department's Prize for Best Undergraduate Engineering Project for developing a SLAM-enabled rover