Harry Dong

harryd@andrew.cmu.edu | www.andrew.cmu.edu/user/harryd/ | www.linkedin.com/in/dongharry | @Real_HDong

Education

Carnegie Mellon University *Pittsburgh, PA*

Electrical & Computer Engineering PhD Candidate *2021 - present*

- Advisor: Prof. Yuejie Chi
- GPA: 4.00
- Anticipated Graduation: 05/26
- **Research interests**: Large Language Model Efficiency, Hardware-aware Algorithms, Sparsity in LLMs, Generation Scaling

UC Berkeley *Berkeley, CA*

Statistics BA & Computer Science BA *2017 - 2021*

- GPA: 3.96 (High Distinction)

Industry Experience

Apple Inc. *Seattle, WA*

AI/ML Intern *May 2024 - Aug 2024*

- Enabled faster generation with Apple Foundation Models on Apple silicon
- Mentor: Tyler Johnson; Manager: Emad Soroush

Air Force Research Lab *Wright-Patterson AFB, OH*

Research Intern *May 2022 - Aug 2022*

- Generative modeling for high-dimensional materials science applications using transformers and diffusion models
- Mentors: Megna Shah & Sean Donegan

Amazon Web Services *Seattle, WA (Remote)*

Software Development Engineer Intern *Jun 2021 - Aug 2021*

- Full stack development of internal service for cloud operations cost modeling
- Received but declined full-time offer to pursue PhD

Amazon Web Services *Seattle, WA (Remote)*

Software Development Engineer Intern *May 2020 - Aug 2020*

- Full stack development of internal services that facilitate server testing for hardware engineers
- Received return offer

Academic Experience

Yuejie Chi Group *Pittsburgh, PA*

Advisor: Prof. Yuejie Chi *Sep 2021 - present*

- Developed a fast, learnable, and provable tensor robust principal component analysis algorithm
- Designing algorithms to improve inference efficiency and scaling in transformers/LLMs

Mobile Sensing Lab *Berkeley, CA*

Advisor: Prof. Alexandre Bayen; Mentor: Theophile Cabannes *May 2019 - May 2021*

- Constructed a model to optimize multi-agent network games with applications in traffic routing
- Explored stochastic controller designs for efficient flow through networks

Lawrence Berkeley National Laboratory & UCSF *Berkeley, CA*

Mentors: Roy Ben-Shalom, Jan Balewski *Jun 2019 - May 2021*

- Improved robustness and model interpretability for prediction of neuron ion conductance properties from voltage responses to stimuli

Honors & Awards

Wei Shen and Xuehong Zhang Presidential Fellowship, 2024
Liang Ji-Dian Graduate Fellowship, 2023
Michel and Kathy Doreau Graduate Fellowship in Electrical and Computer Engineering, 2023
NSF GRFP Honorable Mention, 2023
UC Berkeley High Distinction, 2021

Publications

LLM Inference/Efficiency **Generative AI for Science** **Optimization**

Preprints

ShadowKV: KV Cache in Shadows for High Throughput Long Context LLM Inference
Hanshi Sun, Li-Wen Chang, Wenlei Bao, Size Zheng, Ningxin Zheng, Xin Liu, **Harry Dong**, Yuejie Chi, Beidi Chen
Under review, 2024

Journals

A Lightweight Transformer for Faster and Robust EBSD Data Collection
Harry Dong, Sean Donegan, Megna Shah, Yuejie Chi
Scientific Reports, 2023
- Oral presentation at the *Machine Learning for Scientific Imaging Conference at Electronic Imaging*, 2024.
- Poster presentation at the *Joint Workshop at the Intersection of Materials Science and Machine Learning*, 2023.

Fast and Provable Tensor Robust Principal Component Analysis via Scaled Gradient Descent
Harry Dong, Tian Tong, Cong Ma, Yuejie Chi
Information and Inference, 2023
- Contributed talk at *SIAM MDS22*, 2022.

Conferences

Leveraging Multimodal Diffusion Models to Accelerate Imaging with Side Information
Timofey Efimov, **Harry Dong**, Megna Shah, Jeff Simmons, Sean Donegan, Yuejie Chi
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025
- Presentation/poster at the *Computational Imaging Conference at Electronic Imaging*, 2025.

Prompt-prompted Adaptive Structured Pruning for Efficient LLM Generation
Harry Dong, Beidi Chen, Yuejie Chi
Conference on Language Modeling (COLM), 2024
- Also presented as an **oral** at the *ICML Workshop on Efficient Systems for Foundation Models*, 2024.

Get More with LESS: Synthesizing Recurrence with KV Cache Compression for Efficient LLM Inference
Harry Dong, Xinyu Yang, Zhenyu Zhang, Zhangyang Wang, Yuejie Chi, Beidi Chen
International Conference on Machine Learning (ICML), 2024

Deep Unfolded Tensor Robust PCA with Self-supervised Learning
Harry Dong, Megna Shah, Sean Donegan, Yuejie Chi
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023
- Also presented at the *Third Workshop on Seeking Low-Dimensionality in Deep Neural Networks*, 2023.

Learning Optimal Traffic Routing Behaviors Using Markovian Framework in Microscopic Simulation
Theophile Cabannes, Jiayi Li, Fangyu Wu, **Harry Dong**, Alexandre Bayen
TRB 99th Annual Meeting, 2020

Workshops

Towards Low-bit Communication for Tensor Parallel LLM Inference
Harry Dong, Tyler Johnson, Minsik Cho, Emad Soroush
NeurIPS Workshop on Efficient Natural Language and Speech Processing IV, 2024

Towards Structured Sparsity in Transformers for Efficient Inference
Harry Dong, Beidi Chen, Yuejie Chi
ICML Workshop on Efficient Systems for Foundation Models, 2023

Teaching Experience

CMU 18-786 (Introduction to Deep Learning) *Pittsburgh, PA*

Teaching Assistant & Guest Lecturer *Jan 2024 - May 2024*

- Teaching recitations, maintaining the course website, and hosting office hours for a graduate deep learning class

CMU 18-661 (Introduction to ML for Engineers) *Pittsburgh, PA*

Guest Lecturer *Dec 2022*

- Topics on transformers and their bottlenecks

CMU 18-202 (Mathematical Foundations of Electrical Engineering) *Pittsburgh, PA*

Teaching Assistant *Jan 2022 - May 2022*

- Taught recitations, hosted office hours, and created material (homework and exams) for an undergraduate class

UC Berkeley Student Association of Applied Statistics *Berkeley, CA*

Education Director *Jun 2020 - Dec 2020*

- Led a team of lecturers to teach data science concepts and skills to undergraduates of all levels of expertise

Outreach / Engagement / Service

CMU PhD ECE Student Organization (PESO) *Pittsburgh, PA*

Council Member *Sep 2024 - present*

- Proposing/organizing social and networking events for ECE PhD students at CMU

Faculty Hiring Student Council *Pittsburgh, PA*

Council Member *Jan 2023 - Apr 2023*

- Evaluated CMU ECE faculty candidates' interpersonal relationships between colleagues and students

Cal Ballroom *Berkeley, CA*

Competition Coordinator *May 2019 - May 2020*

- Organized all competition-related events with the Cal Ballroom team
- Publicized events, hired judges, negotiated with other organizations, and hosted competitions with hundreds of participants

Reviewership

- **Journals**: IEEE Transactions on Signal Processing
- **Conferences**: CPAL (2024, 2025)
- **Workshops**: ES-FoMo II (ICML 2024), ENLSP (NeurIPS 2024), FITML (NeurIPS 2024)

Miscellaneous

- **Relevant Coursework**
  - **Math/Statistics**: Theoretical Statistics, Linear Algebra, Stochastic Processes, Time Series, Discrete Math, Real Analysis
  - **Electrical Engineering/Computer Science**: Deep Learning, Algorithms, Convex Optimization, Data Structures, Database Systems, Linear Systems, Adaptive Control
  - **Economics**: Econometrics, Microeconomics, Ethics
- **Programming/Software**: Python, R, MATLAB, Java, SQL, PyTorch, Hugging Face, NumPy, SciPy
- **Languages**: English (native), Mandarin (conversational)
- **Other Activities**: Reading, Dance, Racquetball, Tennis, Cooking
- **Citizenship**: USA