

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

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

JOURNALS

Harry Dong, Sean Donegan, Megna Shah, Yuejie Chi, “A Lightweight Transformer for Faster and Robust EBSD Data Collection,” *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.

Harry Dong, Tian Tong, Cong Ma, Yuejie Chi, “Fast and Provable Tensor Robust Principal Component Analysis via Scaled Gradient Descent,” *Information and Inference*, 2023.

- Contributed talk at *SIAM MDS22*, 2022.

CONFERENCES

Timofey Efimov, **Harry Dong**, Megna Shah, Jeff Simmons, Sean Donegan, Yuejie Chi, “Leveraging Multimodal Diffusion Models to Accelerate Imaging with Side Information,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

- Presentation/poster at the *Computational Imaging Conference at Electronic Imaging*, 2025.

Harry Dong, Beidi Chen, Yuejie Chi, “Prompt-prompted Adaptive Structured Pruning for Efficient LLM Generation,” *Conference on Language Modeling (COLM)*, 2024.

- Also presented as an **oral** at the *ICML Workshop on Efficient Systems for Foundation Models*, 2024.

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

Harry Dong, Megna Shah, Sean Donegan, Yuejie Chi, “Deep Unfolded Tensor Robust PCA with Self-supervised Learning,” *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.

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

WORKSHOPS

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

Harry Dong, Beidi Chen, Yuejie Chi, “Towards Structured Sparsity in Transformers for Efficient Inference,” *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
- **Programming/Software:** Python, R, MATLAB, Java, PyTorch, Hugging Face, NumPy, SciPy
- **Other Activities:** Reading, Dance, Racquetball, Tennis, Cooking