

Harry Dong

✉ hdong920@gmail.com | 🏠 hdong920.github.io | 💻 www.linkedin.com/in/dongharry | 🐦 @Real_HDong

Research vision: I aim to make AI models more *efficient and performant* across domains by leveraging inherent structures and patterns within the architecture, data, features, and hardware. My methods have seen success in NLP, reasoning, and materials science domains. My goal is to maximize AI performance with limited resources.

Research interests: Large Language Model Efficiency, Reasoning, Inference Scaling, Post-training, AI for Science

Education

Carnegie Mellon University

Pittsburgh, PA

ELECTRICAL & COMPUTER ENGINEERING PH.D.

2021 - 2026

- Thesis: Towards Efficient and Performant Generative AI Inference
- Committee: Yuejie Chi (Advisor), Beidi Chen, Andrea Zanette, Bilge Acun

UC Berkeley

Berkeley, CA

STATISTICS BA & COMPUTER SCIENCE BA

2017 - 2021

- GPA: 3.96 (High Distinction)

Honors & Awards

CPAL Rising Stars Award, 2026

Wei Shen and Xuehong Zhang Presidential Fellowship, 2024

Liang Ji-Dian Graduate Fellowship, 2023

Michel and Kathy Doreau Graduate Fellowship, 2023

NSF Graduate Research Fellowships Program Honorable Mention, 2023

UC Berkeley High Distinction, 2021

Industry Experience

Microsoft Research (AI Frontiers)

New York City, NY

SENIOR RESEARCHER

Jun 2026 - present

- Manager: John Langford

Meta (GenAI/Superintelligence Lab)

New York City, NY

RESEARCH SCIENTIST INTERN

May 2025 - Nov 2025

- Designed performative LLM parallel interdependent inference scaling methods for reasoning with reinforcement learning
- Manager: Karthik Abinav Sankararaman

Apple

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 Laboratory

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

SOFTWARE DEVELOPMENT ENGINEER INTERN

Seattle, WA (Remote)

May 2020 - Aug 2020

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

University Experience

Yuejie Chi Group

Pittsburgh, PA

ADVISOR: PROF. YUEJIE CHI

Sep 2021 - May 2026

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

Data-driven Discovery of Optimized Multifunctional Material Systems (D3OM2S)

Pittsburgh, PA

MENTORS: SEAN DONEGAN, MEGNA SHAH, & JEFF SIMMONS

May 2022 - May 2026

- Collaboration with Air Force Research Laboratory to use scalable generative AI in materials science applications

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 interpretability for predicting neuron ion conductance properties from voltage responses to stimuli

Publications

PREPRINTS

- (P1) **Multimodal Diffusion to Mutually Enhance Polarized Light and Low Resolution EBSD Data**
Harry Dong, Timofey Efimov, Megna Shah, Jeff Simmons, Sean Donegan, Marc De Graef, Yuejie Chi
Preprint, 2026

JOURNALS

- (J2) **A Lightweight Transformer for Faster and Robust EBSD Data Collection**
Harry Dong, Sean Donegan, Megna Shah, Yuejie Chi
Scientific Reports, 2023
- **Oral** 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.
- (J1) **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

- (C9) **Generalized Parallel Scaling with Interdependent Generations**
Harry Dong, David Brandfonbrener, Eryk Helenowski, Yun He, Mrinal Kumar, Han Fang, Yuejie Chi, Karthik Abinav Sankararaman
International Conference on Learning Representations (ICLR), 2026
- **Oral & Best Paper Nomination** at *NuerIPS Workshop on Efficient Reasoning*, 2025.
- (C8) **STEM: Scaling Transformers with Embedding Modules**
Ranajoy Sadhukhan, Sheng Cao, **Harry Dong**, Changsheng Zhao, Attiano Purpura-Pontoniore, Yuandong Tian, Zechun Liu, Beidi Chen
International Conference on Learning Representations (ICLR), 2026

- (C7) **Scalable LLM Reasoning Acceleration with Low-rank Distillation**
Harry Dong, Bilge Acun, Beidi Chen, Yuejie Chi
Conference on Parsimony and Learning (CPAL), 2026
 - Early version presented at *ICML Workshop on Long-Context Foundation Models*, 2025.
- (C6) **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
International Conference on Machine Learning (ICML) spotlight, 2025
- (C5) **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.
- (C4) **Prompt-prompted Adaptive Structured Pruning for Efficient LLM Generation**
Harry Dong, Beidi Chen, Yuejie Chi
Conference on Language Modeling (COLM), 2024
 - **Oral** at the *ICML Workshop on Efficient Systems for Foundation Models*, 2024.
- (C3) **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
- (C2) **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.
- (C1) **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

- (W2) **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
- (W1) **Towards Structured Sparsity in Transformers for Efficient Inference**
Harry Dong, Beidi Chen, Yuejie Chi
ICML Workshop on Efficient Systems for Foundation Models, 2023

Talks

- (T9) **Towards Efficient and Performant Generative AI Inference**
Ph.D. Defense, 2026 Pittsburgh, PA
- (T8) **Squeezing the Most Out of LLMs: The Interplay Between Efficiency and Performance**
CPAL Rising Stars Talk, 2026 Tubingen, Germany
- (T7) **Generalized Parallel Scaling with Interdependent Generations**
NeurIPS Workshop on Efficient Reasoning, 2025 San Diego, CA
- (T6) **Leveraging Multimodal Diffusion Models to Accelerate Imaging with Side Information**
ICASSP, 2025 Hyderabad, India
Data-driven Discovery of Optimized Multifunctional Material Systems Funding Review, 2024 Pittsburgh, PA
- (T5) **Prompt-prompted Adaptive Structured Pruning for Efficient LLM Generation**
ICML Workshop on Efficient Systems for Foundation Models, 2024 Vienna, Austria
- (T4) **A Lightweight Transformer for Faster and Robust EBSD Data Collection**
Electronic Imaging, 2024 San Francisco, CA
Data-driven Discovery of Optimized Multifunctional Material Systems Funding Review, 2023 Pittsburgh, PA
- (T3) **Recovering Missing Electron Backscatter Diffraction Microscopy Slices using Transformers**
MIRACLE Forum, 2023 Virtual

- (T2) **Fast and Provable Tensor Robust Principal Component Analysis via Scaled Gradient Descent**
SIAM MDS, 2022 *San Diego, CA*
- (T1) **Learning Missing EBSD Serial Section Slices**
Data-driven Discovery of Optimized Multifunctional Material Systems Funding Review, 2022 *Pittsburgh, PA*

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 - May 2026*
 • Proposing/organizing social and networking events for ECE PhD students at CMU
- Faculty Hiring Student Council** *Pittsburgh, PA*
 COUNCIL MEMBER *2023, 2025*
 • Evaluated CMU ECE faculty candidates' fit with current faculty 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:** ACM Transactions on Intelligent Systems and Technology, IEEE Transactions on Signal Processing
 • **Conferences:** ICLR (2026), CPAL (2024, 2025)
 • **Workshops:** ER (NeurIPS 2025), SCOPE (ICLR 2025), ENLSP (NeurIPS 2024), FITML (NeurIPS 2024), ES-FoMo II (ICML 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, C, SQL, PyTorch, Hugging Face, vLLM, NumPy, SciPy
- **Languages:** English (native), Mandarin (conversational)
- **Other Activities:** Reading, Running, Racquetball, Tennis, Ballroom Dance, Cooking
- **Citizenship:** USA