Wai Tong Chung

Email: wt.chung94@gmail.com Personal Web: waitong94.github.io

Education

Stanford University Stanford, CA
Ph.D. in Mechanical Engineering. Sept. 2018 - June 2024
Advised by Prof. Matthias Ihme in Flow Physics and Computational Engineering Group.
Thesis: *Overcoming Small Datasets in Machine Learning Studies of Multi-physics Flows in Propulsion.* [url]
Research Focus: *AI for Science, High-Performance Computing, Scientific Machine Learning.*

Imperial College London United Kingdom
B.Eng. M.Eng. in Mechanical Engineering *with First Class Honours*. Sept. 2013 - Aug. 2017
Awards: Most Outstanding Thesis (Top 1 of 138 students), Dean's List (Top 10% of 138 students).
Course Focus: *Computational Science, Flow Physics, Statistical Mechanics, Linear Algebra, Probability Theory*.

Experience

Together AI San Francisco, CA
AI Researcher July 2024 - Present
– *Investigating pre- and post-training methods for language models in inference optimization and agentic applications.*
– *Developed speculator model training methods that resulted in the fastest B200 inference of DeepSeek-R1 (July 2025).*
– *Contributed >10K lines of code to a Kubernetes-based project for automating org-wide training/inference experiments.*
– *Authored 1 technical blog on large language model inference optimization through training speculator models.*

Stanford University Stanford, CA
Machine Learning Graduate Research Assistant Sept. 2018 - June 2024
– *Investigated and developed deep learning methods for efficient high-performance computing software in flow physics.*
– *Authored 20+ AI for science and computational engineering refereed publications in top ML and engineering venues.*
– *Contributed significantly to accepted NSF, NASA, U.S. DoE, and Google grant proposals (total worth > $1.5M).*

Lawrence Livermore National Laboratory Livermore, CA
Deep Learning Research Intern June 2022 - Sept. 2022
– *Explored 3D computer vision methods for atmospheric modeling.*
– *Authored 1 publication and 1 conference proceeding in geo-physics venues.*

JPMorgan Chase & Co. United Kingdom
Financial Messaging Software Engineer Sept. 2017 - Aug. 2018
– *Developed, tested, and deployed a Java-based financial messaging application that processed $6T of daily payments.*

Selected Writing

See Google Scholar for full list of academic publications.

Technical Blogs

W.T. Chung, D. Waters, A. May, B. Athiwaratkun. Boosting DeepSeek-R1's Speed with Customized Speculative Decoding. *Together AI*, 2025. [url]

Refereed Journal, Conference, and Workshop Articles

M Ihme†, **W.T. Chung**†. Artificial Intelligence as a Catalyst for Combustion Science and Engineering‡. Accepted in *Proc. Combust. Inst.* 40, 2024. (†*Equal Contribution*. ‡Presented as a plenary lecture at the 40th International Symposium on Combustion, Milan, 2024 [.pdf]

W.T. Chung, B. Akoush, P. Sharma, A. Tamkin, K.S. Jung, J.H. Chen, J. Guo, D. Brouzet, M. Talei, B. Savard, A.Y. Poludnenko, M. Ihme. Turbulence in Focus: Benchmarking Scaling Behavior of 3D Volumetric Super-Resolution with BLASTNet 2.0 Data. *Adv. Neural Inf. Process. Syst. (NeurIPS)* 36, 2023. [.pdf, press]

M. Ihme†, **W.T. Chung**†, A.A. Mishra†. Combustion Machine Learning: Principles, Progress, and Prospects. *Prog. Energy Combust. Sci.* 91:101010, 2022. (†*Equal Contribution*)[.pdf]

W.T. Chung, K.S. Jung, J. H. Chen, M. Ihme. The Bearable Lightness of Big Data: Towards Massive Public Datasets in Scientific Machine Learning. In: *ICML AI4Science Workshop*, 2022. [.pdf]

D.D. Wu, **W.T. Chung**, M. Ihme. ML for Safely Landing on Mars. In: *NeurIPS ML4PS Workshop*, 2022. [.pdf]

W.T. Chung, A.A. Mishra, N. Perakis, M. Ihme. Accelerating High-fidelity Combustion Simulations with Classification Algorithms. In: *AAAI MLPS Spring Symp.*, 2021. [.pdf, video]

W.T. Chung, A.A. Mishra, N. Perakis, M. Ihme. Random Forests for Accelerating Turbulent Combustion Simulations. In: *NeurIPS ML4PS Workshop*, 2020. [.pdf]

Accepted Grant Proposals

NSF Pathways to Enable Open-Source Ecosystems Grant (Awarded $1.2M). PI: M. Ihme, 2024. [info]

Google Award for Inclusion Research Grant (Awarded $60K). PI: M. Ihme, 2022. [info]

US Department of Energy NERSC Grant (Awarded 11.2M core-hours). PI: M. Ihme, 2022. [info]

NASA Early Stage Innovations Grant (Awarded $650K). PI: M. Ihme, 2021. [info]

Honors and Awards

Stanford CS323: *The AI Awakening* **Best Final Project Prize** (Top 4 of 87 Students)	2023
Stanford Human-Centered AI **Affinity Group Award** [info, press]	2023
Stanford Human-Centered AI **Graduate Fellowship** [info, press]	2022-2023
Stanford School of Engineering **Graduate Fellowship**	2018-2019
Imperial College Mechanical Engineering **Most Outstanding Thesis Prize** (Top 1 of 138 Students)	2017
Imperial College Mechanical Engineering **Dean's List** (Top 10% of 138 Students)	2017

Selected Professional Activities

Lead Organizer for *Future Learning Approaches for Modeling and Engineering (FLAME) AI Workshop*, 2023. [info]
Lead Organizer for *Stanford HAI Climate-Centered AI Seminar Series*, 2023. [press]
Affiliate for *Stanford Data Science Center for Open and REproducible Science*, 2023-2024.
Reviewer: *ICLR* 2025; *AISTATS* 2025; *NeurIPS*, 2024, 2025; *ML and the Physical Sciences Workshop at NeurIPS*, 2021, 2022, 2023, 2024; *Synergy of Scientific and Machine Learning Modeling Workshop at ICML*, 2023; *ReScience C (ML Reproducibility Challenge)*, 2023; *AI for Science: Progress and Promises Workshop at NeurIPS*, 2022, 2024; *Proceedings of the Combustion Institute*, 2024; *Signal, Image and Video Processing*, 2024; *ASME Turbomachinery Technical Conference & Exposition*, 2023; *Combustion and Flame*, 2023, 2024; *International Journal of Engine Research*, 2023.

Skills

Programming and Engineering
Proficient: `Python, PyTorch, Slurm, Docker`
Familiar: `C++, C, Kubernetes, MPI, TensorFlow, MATLAB, FORTRAN, Java.`

Languages
Proficient: English, Malay.
Familiar: Mandarin, Cantonese.