Wai Tong Chung

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

Experience

Amazon San Francisco, CA
Member of Technical Staff, AGI Autonomy Dec 2025 - Present
– *Developed continual pre-training, SFT, and RL recipes and infrastructure for agentic large language models.*
– *Achieved 91% on AIME 2026, 68% on SWE-bench Verified and 45% on Terminal-Bench 2.0 with a 100B MoE model.*
– *Extended the internal fork of our Miles-based RL trainer to support continual pre-training and SFT.*
– *Integrated the Miles-based trainer into the team's NVIDIA B200 cluster and maintained it for a team of ∼100.*

Together AI San Francisco, CA
AI Researcher July 2024 - Dec 2025
– *Investigated pre- and post-training methods for inference optimization and agentic language-model applications.*
– *Developed speculator model training methods that resulted in the fastest B200 inference of DeepSeek-R1 (July 2025).*
– *Authored four technical blogs on speculative decoding for reasoning-model inference optimization.*

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.*
– *Wrote accepted NSF, NASA, U.S. DoE, and Google grant proposals (total worth $>$ $1.5M).*

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.*

Education

Stanford University Stanford, CA
Ph.D. in Mechanical Engineering; AI for Science and High-Performance Computing. Sept. 2018 - June 2024

Imperial College London United Kingdom
B.Eng. M.Eng. in Mechanical Engineering, First Class Honours; Best Thesis Prize. Sept. 2013 - Aug. 2017

Selected Writing

See Google Scholar for full list of academic publications.

Technical Blogs

W.T. Chung *et al.* Boosting DeepSeek-R1's Speed with Customized Speculative Decoding. *Together AI Blog*, 2025. [url] J. Wang, **W.T. Chung** *et al.* Together AI Delivers Fastest Inference for the Top Open-Source Models. *Together AI Blog*, 2025. [url,]

S. Zhu, F. Bianchi, **W.T. Chung** *et al.* How Together AI Uses AI Agents to Automate Complex Engineering Tasks: Lessons from Developing Efficient LLM Inference Systems. *Together AI Blog*, 2025. [url]

J. Wang, . . . , **W.T. Chung** *et al.* AdapTive-LeArning Speculator System (ATLAS): A New Paradigm in LLM Inference via Runtime-Learning Accelerators. *Together AI Blog*, 2025.[url, press]

Together AI†. Together AI Delivers Top Speeds for DeepSeek-R1-0528 Inference on NVIDIA Blackwell. *Together AI Blog*, 2025. (†**W.T. Chung** contributed to speculative decoding and benchmarking). [url]

Research Articles

C.-A. Oncescu, Q. Wu, **W.T. Chung** *et al.* Opportunistic Expert Activation: Batch-Aware Expert Routing for Faster Decode Without Retraining. *Int. Conf. Mach. Learn. (ICML)*, 2026. [.pdf]

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 40^{th} International Symposium on Combustion, Milan, 2024) [.pdf]

W.T. Chung *et al.* 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]

Selected Professional Activities

Contributor to `Miles` [github]
Reviewer for *ICLR* 2025–2026; *NeurIPS* 2024–2026; *AISTATS* 2025; and other AI-for-science venues.
Lead Organizer for *Stanford HAI Climate-Centered AI Seminar Series*, 2023. [press]

Skills

Programming and Engineering
Training/Inference: `PyTorch, Miles, Megatron-LM, SGLang, vLLM.`
Systems/Infrastructure: `Kubernetes, Slurm, Ray, MPI.`
Programming: `Python, C++, Java`

Languages

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

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