

William Tong

☎ (509) 990-6102 | ✉ wtong@g.harvard.edu | 🏠 https://wlt.coffee | 📱 wtong98 | 🌐 wlt

Education

Harvard University

Cambridge, MA

PHD IN APPLIED MATH; MS IN DATA SCIENCE | *Kempner Graduate Fellow, GPA: 3.9*

Exp. Graduation: 05/28

Columbia University

New York, NY

BA IN COMPUTER SCIENCE AND STATISTICS | *Summa cum laude, PBK, GPA: 4.0*

Graduation: 05/21

Relevant skills Python, Jax, PyTorch, TensorFlow, LLM engineering, CUDA, SLURM/High Performance Computing, Linux

Experience

AMD - Training and Inference Optimization Team

San Jose, CA

RESEARCH INTERN | ADVISOR: ZHENYU GU

Sep. 2026 - Dec. 2026 (exp)

- My research direction is on agentic harness design for autonomous inference optimization, with focuses on multi-agent scaling dynamics, long context management, and self-improvement limits.

Apple - Machine Learning Research

Cambridge, MA

RESEARCH INTERN | ADVISORS: ERAN MALACH, EMMANUEL ABBE

Feb. 2026 - Sep. 2026

- Developed theory relating gating parameter strength to retrieval performance in state space models, with downstream application to long-context reasoning. Validated results in large-scale neural networks trained for agentic tool use. [Manuscript](#) accepted at *NeurIPS 2026*.
- Proposed novel *weighted state souping* method for agentic tool use with SSMs, and studied the theoretical properties of these models for in-context retrieval in long-context settings. The resulting method boosts long-context performance from failing to near perfect across diverse tool use and retrieval tasks. [Manuscript](#) accepted at *NeurIPS PALM Workshop 2026*.

Pehlevan Group

App. Math Dept., Harvard

GRADUATE RESEARCH ASSISTANT | ADVISOR: CENGIZ PEHLEVAN

Jan. 2022 - present

- Finetuned tens of frontier LLMs on a custom 95 billion token dataset of Lean proofs, studying reasoning traces as they relate to task topology. Developed theoretical analysis to explain results. [Manuscript](#) accepted at *NeurIPS 2026*.
- Developed theory of equality reasoning in multi-layer perceptron (MLP) models, highlighting the understudied importance of learning richness. Validated results on visual reasoning tasks. [Manuscript](#) published in *CCN Proceedings 2025*.
- Measured in-context learning in MLPs, demonstrating that attention mechanisms are not necessary to develop this core ability of large language models. [Manuscript](#) published and presented at *ICLR 2025*.

Murthy Lab / NTT PHI Lab

MCB Dept., Harvard

RESEARCH INTERN | ADVISORS: GAUTAM REDDY, VENKATESH N. MURTHY

Sep. 2021 - August 2023

- Designed end-to-end deep reinforcement learning framework to model trail-following behavior in animals.
- Explored theoretical foundations of automated curriculum learning (ACL) under a teacher-student framework with realistic constraints. [Manuscript](#) published in *PLOS Computational Biology* and presented at *COSYNE 2024*.

Lev, Inc. - Engineering Team

New York, NY

DATA SCIENCE INTERN

May 2021 - Aug. 2021

- Designed, trained, and deployed a large language model to parse term sheets. Model achieved near-perfect accuracy (over 99 percent) on tagging text into target categories (e.g. loan amount, term period, interest rate). Built using PyTorch and deployed on Amazon Web Services (AWS).
- Built email-based queuing system for term sheet ingest using AWS. Queuing service enables asynchronous processing through an email frontend.
- Created React frontend to interact with the trained model, collect human-labeled data, and relabel missed or incorrect text for continual model improvement.

- Designed communication protocol for the OpenTelemetry software telemetry standard. Protocol enables remote, dynamic configuration updates on instrumented applications
- Led adoption of protocol by project maintainers. Protocol accepted as experimental addition to the telemetry standard
- Built reference implementations of the protocol for the Go SDK

Selected publications (6 of 11)

- 2026 **William L. Tong**, Aryo Lotfi, Emmanuel Abbe, Josh Susskind, Cengiz Pehlevan, Eran Malach. Weighted State Souping for Efficient Long-Context Retrieval. *NeurIPS Workshop on Personalized, Aligned, Long-Term Memory for AI Systems*.
- 2026 **William L. Tong**, Aryo Lotfi, Emmanuel Abbe, Kostas Vaggelakos, Vishnu Banna, Etai Littwin, Josh Susskind, Eran Malach. [On the Importance of Gating: Memorization vs. in-context learning in state space models](#). *Conference on Neural Information Processing Systems (NeurIPS)*.
- 2026 **William L. Tong**, Ege Çakar, and Cengiz Pehlevan. [Boule or Baguette? A study on task topology, length generalization, and the benefit of reasoning traces](#). *Conference on Neural Information Processing Systems (NeurIPS)*.
- 2025 **William L. Tong** and Cengiz Pehlevan. [Learning richness modulates equality reasoning in neural networks](#). *Annual Conference on Computational Cognitive Neuroscience (CCN) Proceedings*.
- 2025 **William L. Tong** and Cengiz Pehlevan. [MLPs learn in-context on regression and classification tasks](#). *International Conference on Learning Representations (ICLR)*.
- 2025 **William L. Tong**, Venkatesh Murthy, and Gautam Reddy. [Adaptive algorithms for shaping behavior](#). *PLOS Computational Biology*.

Leadership and teaching

Curr.	Guildmaster	Game Guild. Founded and lead graduate school board game club	<i>Harvard</i>
2026	Res. Advisor	Graduate Residential Life. Led residential floor and organized community events	<i>Harvard</i>
2024	Teaching Fellow	Neural Computation. Graduate course in deep learning theory and neuroscience	<i>Harvard</i>
2023	Teaching Fellow	Machine Learning. Undergraduate introduction to machine learning	<i>Harvard</i>
2021	Teaching Fellow	Deep Learning for NLP. Graduate course in natural language processing	<i>Harvard</i>
2021	Tutor	Stats Help Room. Departmental statistics tutor	<i>Columbia</i>
2021	Teaching Assist.	Machine Learning. Graduate introduction to machine learning	<i>Columbia</i>
2021	Curriculum dev.	Global Youth Mentorship Initiative. Developed teaching material and organized socials	<i>Columbia</i>
2020	Teaching Assist.	Clean object-oriented design. Graduate course in software design practices	<i>Columbia</i>
2019	Teaching Assist.	Hon. intro to CS. Undergraduate honors-level introduction to computer science	<i>Columbia</i>
2018	Instructor	Columbia Splash. Designed and delivered introductory stats lecture	<i>Columbia</i>
2017	Instructor	Ruzhou Bureau of Education. Taught English lessons through outreach program	<i>China</i>