James Liu

jamesliu2004@gmail.com | Linkedin | jameszliu.com

Education

Massachusetts Institute of Technology Cambridge, MA

Candidate for B.S. in EECS: Artificial Intelligence and Decision Making — GPA: 5.0/5.0 *2022-2025*

- **Graduate Coursework**: Information and Inference [6.7800], Advances in Computer Vision [6.8300], Machine Learning [6.7900].
- **Relevant Coursework**: Design and Analysis of Algorithms [6.1220], Probability and Random Variables [18.600], Natural Language Processing [6.861], Statistics [18.650], Signal Processing [6.3000], Representation + Inference + Reasoning in AI [6.4110], Networks [6.3260], Foundations of Cryptography.
- **Associations**: Research Science Institute (RSI), Phi Kappa Theta Fraternity, Startlabs, MIT AI Alignment, AI@MIT , Neo Scholar Finalist.
- **Awards**: USA Computing Olympiad (USACO) Platinum Division, USA Physics Olympiad (USAPhO) Semifinalist, National Security Agency (NSA) Special Award in Cybersecurity.

Experience

Anthropic 06/2025 – Present

Member of Technical Staff | Performance Engineering *San Francisco, CA*

- Building out the TPU kernel stack.

Liquid AI 01/2025 – 04/2025

Member of Technical Staff | Pretraining *Boston, MA*

- Built pre-training infrastructure for memory efficient Mixture-of-Expert models.
- Developed key CUDA kernels for proprietary Liquid Foundation Models (LFMs).
- Maintained and upgraded the GPU inference stack.

Together AI 06/2024 – 09/2024

Machine Learning Researcher | Turbo Team *San Francisco, CA*

- Developed methodology to induce 40-50% activation sparsity in modern LLMs with minimal degradation.
- Created a hardware-aware CUDA kernel for sparse matrix-vector multiplication, realizing a 1.53-1.8× speedup in single-batch edge inference.
- Resulted in a first-author paper, and will present at ICLR 2025.

Massachusetts Institute of Technology 10/2023 – 02/2024

Undergraduate Researcher | HAN Lab *Cambridge, MA*

- Worked on attention-level explanation for the "Lost in the Middle" phenomenon.
- Developed methodology to efficiently compress fine-tune weight information in Large Language Models (LLMs) to 1 bit, reducing GPU memory requirements by $> 10\times$, and per-user generation latency by $> 10\times$ in multi-tenant settings.
- Resulted in first author paper, and presented at NeurIPS 2024.

Publications

Training-Free Activation Sparsity in Large Language Models 08/2024

***James Liu**, Pragaash Ponnusamy, Tianle Cai, Han Guo, Yoon Kim, Ben Athiwaratkun* *ICLR 2025 (Spotlight)*

BitDelta: Your Fine-Tune May Only Be Worth One Bit 02/2024

James Liu, Guangxuan Xiao, Kai Li, Jason D. Lee, Song Han, Tri Dao, Tianle Cai** *NeurIPS 2024 (Poster)*

Projects

SGLang | *Pytorch* 03/2025 – Present

- Contributor to SGLang (12k+ stars), an open-source LLM inference engine adopted by xAI, AMD, NVIDIA, and more.
- Integrated the EAGLE-3 speculative decoding method, speeding up low-batch inference by $> 30\%$.

Optimizing RecurrentGemma | *Python, Pytorch, Triton* 11/2024

- Improved the inference speed of Google's RecurrentGemma-2B Pytorch implementation by $> 6\times$ with no model degradation, using a combination of kernel-level and systems-level optimizations.
- Refactored codebase to be compatible with `torch.compile` during decode.
- Created custom hardware-aware triton kernel, speeding up the sequential scan during prefill by $> 70\times$.

Skills & Interests

Languages: Python, C/C++, CUDA, LaTeX
Frameworks: Pytorch, Transformers, React, FastAPI
Developer Tools: Git, Amazon Web Services (AWS), Google Cloud Platform (GCP), HuggingFace
Libraries: Triton, NumPy, Matplotlib
Interests: Artificial Intelligence, Entrepreneurship, Hip-Hop, League of Legends (Diamond 3), Skiing