

Ying Feng

yinggggfgeng.github.io | yingfeng@mit.edu | [+1 \(878\) 670-6762](tel:+18786706762)

EDUCATION

Massachusetts Institute of Technology
Ph.D. in Computer Science · Advisor: [Piotr Indyk](#)

Sep 2024 – May 2028 (expected)
Cambridge, MA

Carnegie Mellon University
B.S. in Computer Science, with University Honors

Sep 2020 – May 2024
Pittsburgh, PA

RESEARCH HIGHLIGHTS

My research develops provably efficient methods for LLM inference and retrieval.

Fast Quantization for Inference and Retrieval

- Proved that randomized Hadamard transforms preserve the error guarantees of rotation-based quantizers (e.g., EDEN, TurboQuant) while reducing rotation time from $\Theta(d^2)$ to $\Theta(d \log d)$. [2]
- Implemented Hadamard-based quantizers that nearly match the original quantizers' nearest-neighbor recall on 1536- and 3072-dim OpenAI embeddings while running 2.6–4.6× faster.

KV-Cache Compression

- Proved nearly matching upper and lower bounds on the KV-cache memory required to approximate attention during autoregressive decoding. [1]
- Showed that under the small key–query similarities measured in Llama 3.1, Qwen3, and Ministral 3, uniform token eviction is provably near-optimal, explaining why simple sampling baselines remain competitive. [11]

Vector Search and Multi-Vector Retrieval

- Built an approximate nearest-neighbor index that answers every query correctly in $d \cdot \text{polylog}(n)$ time; powered by our new fast embeddings with the smallest known output dimension. [4, 12]
- Designed algorithms for multi-vector retrieval under the Chamfer similarity (ColBERT-style scoring), attaining near-optimal runtimes for underlying ℓ_p metrics, $p \in [1, 2]$. [3, 6]

Adversarially Robust Search and Sketching

- Designed vector search, regression, and frequency-estimation algorithms that stay correct under adversarial queries, using differential privacy and cryptographic techniques. [5, 7, 9]

PUBLICATIONS AND MANUSCRIPTS

*Authors are listed alphabetically; exceptions use * to indicate first author(s).*

Conference Publications

1. **Towards Tight Bounds for Streaming Attention**
Justin Y. Chen, **Ying Feng**, Piotr Indyk, Michael Kapralov, Ekaterina Kochetkova, and Boris Prokhorov
SODA 2027
2. **Provable Quantization with Randomized Hadamard Transform**
Ying Feng, Piotr Indyk, Michael Kapralov, Dmitry Krachun, and Boris Prokhorov
NeurIPS 2026
3. **Fast Approximate L_p Chamfer Distance via Lopsided Embeddings and Structured JL**
Ying Feng and David P. Woodruff
NeurIPS 2026
4. **Fast and Compact Random Mappings with Uniform Guarantees and Applications**
Ying Feng and Piotr Indyk
STOC 2026

5. **On Differential Privacy for Adaptively Solving Search Problems via Sketching**
Shiyuan Feng, **Ying Feng**, George Z. Li, Zhao Song, David P. Woodruff, and Lichen Zhang
ICML 2025. Oral presentation (top 1% of submissions).
6. **Even Faster Algorithm for the Chamfer Distance**
Ying Feng and Piotr Indyk
ICALP 2025
7. **Fast White-Box Adversarial Streaming Without a Random Oracle**
Ying Feng, Aayush Jain, and David P. Woodruff
ICML 2024
8. **A Real-Time Rescheduling Algorithm for Multi-Robot Plan Execution**
Ying Feng*, Adittyo Paul, Zhe Chen, and Jiaoyang Li
ICAPS 2024
9. **Improved Algorithms for White-Box Adversarial Streams**
Ying Feng and David P. Woodruff
ICML 2023
10. **A Fast Rescheduling Algorithm for Real-Time Multi-Robot Coordination**
Adittyo Paul*, **Ying Feng***, and Jiaoyang Li
SoCS 2023

Under Review

11. **Attention under Bounded Key-Query Similarity: Space Complexity and Transformer Anisotropy**
Justin Y. Chen, **Ying Feng**, Piotr Indyk, Michael Kapralov, Ekaterina Kochetkova, and Boris Prokhorov
12. **Ultra-Fast Deterministic Approximate Near Neighbor Search in High Dimensions**
Ying Feng and Piotr Indyk
13. **Condition Number-Independent Sparse Linear Regression on Random Supports**
Ying Feng

Technical Report

- **Accelerating Scientific Research with Gemini: Case Studies and Common Techniques**
David P. Woodruff*, et al. (including **Ying Feng**)
arXiv preprint, 2026

SERVICE

Conference reviewer: NeurIPS (2025, 2026), ICML (2026, **Gold Reviewer**), ICLR (2026), SODA (2026), ITCS (2026, 2027), RANDOM (2025).

Journal reviewer: SIAM Journal on Mathematics of Data Science (SIMODS).

AWARDS

- | | |
|--------------------------------------|-------------|
| • MathWorks Fellowship | 2026 – 2027 |
| • MIT Akamai Presidential Fellowship | 2024 – 2025 |

TEACHING

Teaching Assistant, Matrix Algebra with Applications	2023
Carnegie Mellon University	Pittsburgh, PA

SKILLS

Programming: Python, C/C++.

ML & Systems: PyTorch, Hugging Face Transformers, FAISS, Slurm, Git, CUDA (basic).