Avner May

Brooklyn, NY • avnermay@gmail.com • (301)518-5058
www.linkedin.com/in/avnermay/ • avnermay.github.io

Work Experience

Together.ai — **New York, NY**
Staff Research Scientist — ***Sep. 2023 - present***
Working on making training and inference for large language models faster and more memory efficient.

Google Speech Recognition Group — **New York, NY**
Research Scientist — ***Oct. 2020 - Sep. 2023***
Worked on self-supervised learning for audio-visual speech recognition models. The goal was to use untranscribed videos of people speaking (e.g., audio + lip movements) to improve the performance of audio-visual speech recognition systems.

Google Research – Large Scale Machine Learning Research Group — **New York, NY**
Research Intern — ***Summer 2015***
Worked on model compression, a research area which attempts to train more compact models in the case where larger more powerful models already exist. Performed experiments using Torch.

Microsoft Research – Speech and Dialogue Research Group — **Redmond, WA**
Research Intern — ***Summer 2014***
Worked on training acoustic models from the raw speech signal. Specifically, was interested in seeing whether it was possible to train the matrices which perform the Fourier transform and mel-binning within the classic log-mel feature acoustic frontend.

Microsoft Corporation – Windows Communication Foundation (WCF) — **Redmond, WA**
Software Development Engineer — ***Aug. 2009 - Jul. 2011***
Designed and implemented features to facilitate the development of distributed applications.
Honors: Received "Gold Star Bonus Award" for contributions to the team.

Microsoft Corporation – Windows Workflow Foundation (WF) — **Redmond, WA**
Software Development Engineer Intern — ***Summer 2008***
Designed and implemented a program for validating Windows Workflow programs. Integrated it with Microsoft Visual Studio.

Education

Stanford University — **Stanford, CA**
Postdoctoral Scholar — ***Jan. 2018 - Jul. 2020***
Advisor: Christopher Ré

Columbia University — **New York, NY**
MS/PhD in Computer Science — ***Sep. 2011 - Dec. 2017***
GPA: 4.07/4.00
Advisor: Michael Collins
Honors: Recipient of the Department Chair's Distinguished Fellowship

Harvard University — **Cambridge, MA**
Bachelor of Arts in Mathematics, Secondary Field Computer Science — ***Jun. 2009***
GPA: 3.60/4.00
Honors: Certificate of Distinction in Teaching (Spring 2008).

PUBLICATIONS

Minions: Cost-efficient Collaboration Between On-device and Cloud Language Models
A. Narayan*, D. Biderman*, S. Eyuboglu*, **A. May**, S. Linderman, J. Zou, C. Ré. ICML 2025.

MagicDec: Breaking the Latency-Throughput Tradeoff for Long Context Generation with Speculative Decoding
R. Sadhukhan*, J. Chen*, Z. Chen, V. Tiwari, R. Lai, J. Shi, I. Yen, **A. May**, T. Chen, B. Chen. ICLR 2025.

The Mamba in the Llama: Distilling and Accelerating Hybrid Models
J. Wang*, D. Paliotta*, **A. May**, A. Rush, T. Dao. NeurIPS 2024.

SpecExec: Massively Parallel Speculative Decoding for Interactive LLM Inference on Consumer Devices
R. Svirschevski*, **A. May***, Z. Chen*, B. Chen, Z. Jia, M. Ryabinin. NeurIPS 2024.

Sequoia: Scalable, Robust, and Hardware-Aware Speculative Decoding
Z. Chen*, **A. May***, R. Svirschevski*, Y. Huang, M. Ryabinin, Z. Jia, B. Chen. NeurIPS 2024.

Contextual Embeddings: When are they worth it?
S. Arora*, **A. May***, J. Zhang, C. Ré. ACL 2020.

Understanding the Downstream Instability of Word Embeddings.
M. Leszczynski, **A. May**, J. Zhang, S. Wu, C. Aberger, C. Ré. MLSys 2020.

On the Downstream Performance of Compressed Word Embeddings.
A. May, J. Zhang, T. Dao, C. Ré. NeurIPS 2019 (Spotlight, 3% acceptance).

Low-Precision Random Fourier Features for Memory Constrained Kernel Approximation.
J. Zhang*, **A. May***, T. Dao, C. Ré. AISTATS 2019.

Kernel Approximation Methods for Speech Recognition.
A. May, A.B. Garakani, Z. Lu, D. Guo, K. Liu, A. Bellet, L. Fan, M. Collins, D. Hsu, B. Kingsbury, M. Picheny, F. Sha. JMLR 2019.

Compact Kernel Models for Acoustic Modeling via Random Feature Selection.
A. May, M. Collins, D. Hsu, B. Kingsbury. ICASSP 2016.

A Comparison Between Deep Neural Nets and Kernel Acoustic Models for Speech Recognition.
Z. Lu, D. Guo, A.B. Garakani, K. Liu, **A. May**, A. Bellet, L. Fan, M. Collins, B. Kingsbury, M. Picheny, F. Sha. ICASSP 2016.

Filter & follow: How social media foster content curation.
A. May, A. Chaintreau, N. Korula, S. Lattanzi. SIGMETRICS 2014.

COMMUNITY SERVICE

Reviewer for ICLR 2018, 2022-2024, ICML 2017-2020, 2022 (2019 Top Reviewer), NeurIPS 2017-2019, 2022-2023, IJCAI 2019-2020 (2019 Distinguished PC member), AAAI 2020, 2022, ICASSP 2023 (Outstanding Reviewer), EMNLP 2022, JMLR, IEEE Transactions on Multimedia.

SKILLS

Computer: Python, PyTorch, Tensorflow, Matlab, Java, C#, Linux, C, C++, CUDA, AWS, Apache Beam.
Language: ***Spanish***: Native speaker. ***Hebrew***: Proficient.