JACOB KAHN

PERSONAL INFORMATION

Email	jacobkahn1@gmail.com
Website	jacobkahn.me
GitHub	github.com/jacobkahn
Scholar	Google Scholar

Research Manager and Engineer at FAIR, Meta AI, developing scalable systems and models for reasoning, code generation, and multimodal learning.

Research Areas: Deep learning, distributed systems, reasoning, code generation, multimodal learning, and speech.

EXPERIENCE

2018 - present

Research Manager and Engineer, FAIR, Meta AI
Menlo Park, California; New York City, New York.
Technical leader across reasoning, code generation, and multimodal modeling, working full-stack on model–system co-design for scaling deep learning. Recruited and mentored researchers to senior and staff roles at Meta.

- Reasoning and Code Generation: Co-leading Code World Model (CWM). Direction for scaling, data generation, RL infrastructure, and inference; leading a team of scientists and engineers; steering over 30 ExaFLOPs of compute.
- Multimodal Modeling: Led inference, infrastructure, and scaling for Chameleon and contributed to Transfusion. Drove projects in end-to-end speech recognition, including Libri-Light and wav2letter.
- Open-Source Frameworks: Led development of Flashlight, co-led Shumai and wav2letter, projects that influenced PyTorch `distributed`, `compile`, and Meta's ML infrastructure roadmaps.
- Hardware and scaling strategy: Planning for Meta's GPU clusters and shaping hardware strategy for large-model scaling; guiding multi-billion-dollar compute investments.

2024 - present

Computer Science Faculty, University of Pennsylvania
Philadelphia, Pennsylvania.
Affiliated with Penn's NetDB Laboratory.
Design and teach *CIS 5690: GPU Programming and Machine Learning Systems*.

2016 - 2018

Engineering Intern, Facebook
Menlo Park, California.
Built large, globally distributed systems and algorithms for high-performance stream processing in C++.

EDUCATION

2016 - 2018

M.S.E. in Computer and Information Science, University of Pennsylvania.

2014 - 2018

Jerome Fisher Program in Management and Technology
University of Pennsylvania

B.S.E. in Computer and Information Science, Penn Engineering
Thesis: *Computer Vision for Multiplayer Anchoring in Real-Time Augmented Reality Systems*

B.S. in Economics with concentrations in Operations Research and Management, The Wharton School

Selected Publications

Full list available on Google Scholar.

2025

CWM: An Open-Weights LLM for Research on Code Generation with World Models
Code World Model Team, **(Core Team, Frameworks & Infrastructure Lead)**
arXiv Preprint 2510.02387

Efficient Hardware Scaling and Diminishing Returns in Large-Scale Training of Language Models
Jared Fernandez, Luca Wehrstedt, Leonid Shamis, Mostafa Elhoushi, Kalyan Saladi, Yonatan Bisk, Emma Strubell, **Jacob Kahn**
In *Transactions on Machine Learning Research, 2025*

Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model
Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, **Jacob Kahn**, Xuezhe Ma, Luke Zettlemoyer, Omer Levy
In *International Conference on Learning Representations (Oral)* Singapore, 2025

2024

Chameleon: Mixed-modal Early-Fusion Foundation Models
Chameleon Team **(Evaluation & Inference Lead)**
arXiv Preprint 2405.09818

2023

Reasoning over Public and Private Data in Retrieval-Based Systems
Simran Arora, Patrick Lewis, Angela Fan, **Jacob Kahn***, Christopher Ré*
in the *Transactions of the Association for Computational Linguistics (TACL), presented at the Proceedings of the ACL, Toronto, Canada, 2023.*
also in *AAAI — workshop on Knowledge Augmented Methods for NLP (Oral), Washington D.C. 2023*
* = Equal advising.

RA-DIT: Retrieval-Augmented Dual Instruction Tuning
X. Lin, X. Chen, M. Chen, W. Shi, M. Lomeli, R. James, P. Rodriguez, **J. Kahn**, G. Szilvasu, M. Lewis, L. Zettlemoyer, S. Yih
In *International Conference on Learning Representations (ICLR), Vienna, Austria, 2024*

2022

Flashlight: Enabling Innovation in Tools for Machine Learning
J. Kahn, V. Pratap, T. Likhomanenko, Q. Xu, A. Hannun, J. Cai, P. Tomasello, A. Lee, E. Grave, G. Avidov, B. Steiner, V. Liptchinsky, G. Synnaeve, R. Collobert
In *International Conference on Machine Learning (ICML) (Spotlight), Baltimore, Maryland, 2022*

2021

slimIPL: Language-Model-Free Iterative Pseudo-Labeling
Tatiana Likhomanenko*, Qiantong Xu*, **Jacob Kahn**, Gabriel Synnaeve, Ronan Collobert
In *Proceedings of Interspeech, Brno, Czech Republic, 2021*
* = Equal contribution.

2020

Self-Training for End-to-End Speech Recognition
Jacob Kahn, Ann Lee, Awni Hannun
In *Proc. of the 45th IEEE International Conference in Acoustic, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020.*

Libri-Light: A Benchmark for ASR with Limited or No Supervision
J. Kahn*, M. Rivière*, W. Zheng*, E. Kharitonov*, Q. Xu*, P.E. Mazaré*, J. Karadayi*, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, E. Dupoux
In *Proc. of the 45th IEEE International Conference in Acoustic, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020.*
* = Equal contribution.

End-to-End ASR: from Supervised to Semi-Supervised Learning with Modern Architectures
G. Synnaeve*, Q. Xu*, **J. Kahn***, E. Grave*, T. Likhomanenko*, V. Pratap, A.

Sriram, V. Liptchinsky, R. Collobert
In *ICML Workshop on Self-Supervision in Audio and Speech, Virtual, 2020*
* = Equal contribution.

Differentiable Weighted Finite-State Transducers
Awni Hannun, Vineel Pratap, **Jacob Kahn**, Wei-Ning Hsu
arXiv Preprint 2010.01003

2019 wav2letter++: A Fast Open-source Speech Recognition System
Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, **Jacob Kahn**, Gabriel Synnaeve, Vitaliy Liptchinsky, Ronan Collobert
In *The 44th IEEE International Conference in Acoustic, Speech and Signal Processing (ICASSP), Brighton, UK, 2019.*

invited talks

2025 *Tracing Program Execution for Neural Planning and Reasoning*
AI Engineer, New York, NY, scheduled for November 2025

Scaling Asynchronous RL Recipes for Learning to Code
Meta Superintelligence Labs, St. Louis, MO, scheduled for November 2025

2024 *Trends in Deep Learning Computation and Frameworks*
Intel, Santa Clara, CA, September 2024

Multimodal Generative Models: Chameleon and Beyond
Charles River Ventures and Greylock, San Francisco, CA, August 2024

GPU Computing at Scale
Guest Lecture, CIS 565: GPU Programming & Architecture, Penn, Philadelphia, PA, July 2024

2023 *Shumai: Fast, Flexible Machine Learning in TypeScript*
Google DeepMind, Mountain View, CA, April 2023

2020 *Scaling Deep Learning for Automatic Speech Recognition*
NVIDIA GPU Technology Conference, San Jose, CA, March 2020

teaching and service

University of Pennsylvania GPU Computing for Machine Learning Systems · CIS 5690 · Instructor
Created and teach a graduate course on GPU programming with CUDA, large-scale deep learning systems, performance engineering, and compilers. Project-based, in CUDA and C++.

Head TA (as student): Algorithms (CIS 320), Graduate Operating Systems (CIS 548)
TA Experience: Undergraduate and graduate courses in distributed systems, computer architecture, and databases, all in C/C++.

Reviewing Regular reviewer at NeurIPS, ICLR, and ICML.