E. Kelly Buchanan

kelly.buchanan@stanford.edu · ekbuchanan.com · github.com/ekellbuch · linkedin.com/in/ekellbuch

Education

Columbia University — New York, NY
Ph.D in Neurobiology and Behavior — 2018 – 2024
M.S. in Neurobiology and Behavior
Advisors: Liam Paninski and John Cunningham
Thesis: Building reliable machine learning systems for neuroscience.

University of Kansas — Lawrence, KS
M.S. Electrical Engineering with *honors* — 2017
B.S. Electrical Engineering, minor in Mathematics
Advisors: Randolph Nudo and Yang Yi
Thesis: Methods to study temporal coding in biological and artificial networks.

Research & Professional Experience

Stanford University – Postdoctoral Fellow — Palo Alto, CA, U.S.
Advisors: Scott Linderman and Christopher Ré — 07/24 –

- Developed Weaver, a weak-verifier aggregation method that narrows the generation–verification gap while minimizing human annotation; adopted in PyTorch's scalable RL pipeline.
- Terminal-Bench, a benchmark for evaluating autonomous LLM-based agents on realistic command-line tasks; and Harbor, a containerized multi-scaffold framework for agent evaluation and post-training.
- Proposed a unifying framework for parallel inference in nonlinear sequence models on memory-dense hardware.

Google AI – Ph.D. Student Researcher — New York, NY, U.S.
Group: Reliable Deep Learning. Mentors: Dustin Tran and Kevin P. Murphy — 02/22 – 10/22

- Improved robustness and uncertainty quantification across tasks, modalities, and scales by developing drop-in architecture extensions (Plex) for large language and vision models up to 1B parameters.

Google X – Ph.D. Resident in AI — New York, NY, U.S.
Group: Rapid Evaluation. Lead: Benoit Schillings — 05/21 – 09/21

- Evaluated transformer alternatives for scientific time series and video data; benchmarked for efficiency and scalability. Received Innovation Award.

Columbia University – Research scientist — New York, NY, U.S.
Center for Theoretical Neuroscience. Advisor: Liam Paninski — 08/17 – 07/18

- Built machine learning systems to process video and time series data, including for neural data denoising and demixing, semi-supervised animal pose estimation, and unsupervised behavioral segmentation.

KU Medical Center – Researcher — Kansas City, KS, U.S.
Cortical Plasticity Lab. Advisor: Randolph Nudo — 2014 – 2017

- Developed algorithms for automated spike detection and sorting for in vivo extracellular recordings.

KU Brain-Inspired Computing Lab – Visiting Researcher — Rome, NY, U.S.
Advisor: Yang Yi. Investigated delay-coupled reservoirs for nonlinear processing of binary time series. — Summer 2016

Italian Institute of Technology, Rehab Technologies Lab – Visiting Researcher — Genoa, Italy
Advisor: Michela Chiappalone. Developed Spike detection algorithms for closed-loop cortical stimulation. — 2016

IBM, Power Systems Qualification – Software Engineering Intern — Rochester, MN, U.S.
Automated test and validation infrastructure for enterprise storage hardware. — 2012 – 2013

Publications

∗ indicates equal contribution

[1] Catching Benchmark Rot: A Systematic Audit of Terminal-Bench 2.0
E. Kelly Buchanan, Scott Linderman, Christopher Ré. *Under Review, 2026.*

[2] Hierarchical Text Generation with Next-Scale Language Modeling (NSM)
Hermann Kumbong, Ishan S. Khare, Andrew Woen, Charles Ding, **E. Kelly Buchanan**, Simran Arora, Azalia Mirhoseini, Christopher Ré. *Under Review, 2026.*

[3] Harbor Index: Infrastructure and a Curated Meta-Dataset for Large-Scale Agentic Evaluation
Harbor Team. *Under Review, 2026.*

[4] OpenThoughts-Agent: Data Recipes for Agentic Models
OpenThoughts Team. *Under Review, 2026.*

[5] Test-Time Scaling Makes Overtraining Compute-Optimal
Nicholas Roberts, Sungjun Cho, Zhiqi Gao, Tzu-Heng Huang, Albert Wu, Gabriel Orlanski, Avi Trost, **E. Kelly Buchanan**, Aws Albarghouthi, Frederic Sala. *arXiv:2604.01411, 2026.*

[6] Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces
Mike Merrill*, Alexander Glenn Shaw*, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, **E. Kelly Buchanan**, et al. *ICLR, 2026.*

[7] A Unifying Framework for Parallelizing Sequential Models with Linear Dynamical Systems
Xavier Gonzalez*, **E. Kelly Buchanan***, Hyun Dong Lee, Jerry Weihong Liu, Ke Alexander Wang, David M. Zoltowski, Christopher Ré, Scott Linderman. *TMLR, 2026.*

[8] Shrinking the Generation-Verification Gap with Weak Verifiers
Jon Saad-Falcon*, **E. Kelly Buchanan***, Mayee Chen*, Tzu-Heng Huang, Brendan McLaughlin, Tanvir Bhathal, Shang Zhu, Ben Athiwaratkun, Frederic Sala, Scott Linderman, Azalia Mirhoseini, Christopher Ré. *NeurIPS, 2025.*

[9] Archon: An architecture search framework for inference-time techniques
Jon Saad-Falcon, Adrian Gamarra, Shlok Natarajan, Nahum Maru, Hristo Todorov, Etash Guha, **E. Kelly Buchanan**, Mayee Chen, Neel Guha, Christopher Ré, Azalia Mirhoseini. *ICML, 2025.*

[10] Pathologies of Predictive Diversity in Deep Ensembles
Taiga Abe, **E. Kelly Buchanan**, Geoff Pleiss, John P. Cunningham. *TMLR, 2024.* **Featured Certification (top 20% of submissions)**.

[11] The Effects of Ensembling on Long-Tailed Data
E. Kelly Buchanan, Geoff Pleiss, Yixin Wang, John P. Cunningham. *NeurIPS, 2023, Heavy Tails in ML Workshop*.

[12] Reliability Benchmarks for Semantic Segmentation
E. Kelly Buchanan, Michael W. Dusenberry, Jie Ren, Kevin P. Murphy, Balaji Lakshminarayanan, Dustin Tran. *NeurIPS, 2022, Workshop on Distribution Shifts*.

[13] The Best Deep Ensembles Sacrifice Predictive Diversity
Taiga Abe*, **E. Kelly Buchanan***, Geoff Pleiss, John Cunningham. *NeurIPS, 2022, I Can't Believe It's Not Better Workshop*. **Entropic Award for Most Surprising Negative Result**.

[14] Deep ensembles work, but are they necessary?
Taiga Abe*, **E. Kelly Buchanan***, Geoff Pleiss, Richard Zemel, John Cunningham. *NeurIPS, 2022.*

[15] Plex: Towards Reliability using Pretrained Large Model Extensions
Dustin Tran, Jeremiah Liu, Mike Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G.J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, **E. Kelly Buchanan**, Kevin P. Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, Balaji Lakshminarayanan. *ICML, 2022, Pre-training Workshop*. **Contributed talk (top 5% of accepted papers)**.

AI for Neuroscience Publications

[16] Extracting task-relevant preserved dynamics from contrastive aligned neural recordings
Yiqi Jiang*, Kaiwen Sheng*, Yujia Gao, **E. Kelly Buchanan**, Yu Shikano, Seung Je Woo, Yixiu Zhao, Tony Hyun Kim, Fatih Dinc, Scott Linderman, Mark Schnitzer. *NeurIPS, 2025.* **(Spotlight)**.

[17] Brain-wide representations of prior information in mouse decision-making
International Brain Laboratory. *Nature, 2025*.

[18] Reproducibility of in vivo electrophysiological measurements in mice
International Brain Laboratory. *eLife, 2025*.

[19] Batik: behavior discovery, interpretation and annotation directly from raw video using video LLMs
Aditya Nair, Rohan Kolhe, Nestor Coria, Jadon Hale, Jineun Kim, Angel Wang, Amit Vinograd, Dan Biderman, **E. Kelly Buchanan**, Pietro Perona, Scott Linderman. *Under Review, Nature Methods*.

[20] Brain-to-Text Benchmark'24: Lessons Learned
Francis R. Willett, Jingyuan Li, Trung Le, Chaofei Fan, Mingfei Chen, Eli Shlizerman, Yue Chen, Xin Zheng, Tatsuo S. Okubo, Tyler Benster, Hyun Dong Lee, Maxwell Kounga, **E. Kelly Buchanan**, David Zoltowski, Scott W. Linderman, Jaimie M. Henderson. *ArXiv, 2024*.

[21] Neuroscience cloud analysis as a service
Taiga Abe, Ian Kinsella, Shreya Saxena, **E. Kelly Buchanan**, Joao Couto, John Briggs, Sian Kitt, Ryan Glassman, John Zhou, Liam Paninski, John P. Cunningham. *Neuron, 2022*.

[22] Semi-supervised sequence modeling for improved behavioral segmentation
Matthew R. Whiteway, Evan S. Schaffer, Anqi Wu, **E. Kelly Buchanan**, Omer F. Onder, Neeli Mishra, Liam Paninski. *CVPR, 2021, Computer Vision for Animal Behavior Workshop*.

[23] Partitioning variability in behavioral videos using semi-supervised deep generative models
Matthew Whiteway, Daniel Biderman, **E. Kelly Buchanan**, Anqi Wu, Mario Dipoppa, Yoni Friedman, International Brain Laboratory, John Cunningham, Liam Paninski. *PLoS Computational Biology, 2021*.

[24] DeepGraphPose: a semi-supervised deep graphical model for improved animal pose tracking
Anqi Wu*, **E. Kelly Buchanan***, Matthew Whiteway, Michael Schartner, Guido Meijer, Jean-Paul Noel, Erica Rodriguez, Claire Everett, Amy Norovich, Evan Schaffer, Neeli Mishra, C. Daniel Salzman, Dora Angelaki, Andres Bendesky, John P. Cunningham, Liam Paninski. *NeurIPS, 2020*.

[25] Voltage imaging and optogenetics reveal behaviour-dependent changes in hippocampal dynamics
Yoav Adam, Jeong Kim, Shan Lou, Yongxin Zhao, Michael Xie, Daan Brinks, Hao Wu, Mohammed Mostajo-Radji, Simon Kheifets, Vicente Parot, Selmaan Chettih, Katherine Williams, Benjamin Gmeiner, Samouil Farhi, Linda Madisen, **E. Kelly Buchanan**, Ian Kinsella, Ding Zhou, Liam Paninski, Christopher Harvey, Hongkui Zeng, Paola Arlotta, Robert Campbell, Adam Cohen. *Nature, 2019*.

[26] Penalized matrix decomposition for denoising, compression, and demixing of functional imaging data
E. Kelly Buchanan*, Ian Kinsella*, Ding Zhou*, Rong Zhu, Pengcheng Zhou, Felipe Gerhard, John Ferrante, Ying Ma, Sharon Kim, Mohammed Shaik, Yajie Liang, Rongwen Lu, Jacob Reimer, Paul Fahey, Taliah Muhammad, Graham Dempsey, Elizabeth Hillman, Na Ji, Andreas Tolias, Liam Paninski. *BioRxiv, 2019*.

[27] Quantifying the behavioral dynamics of C. elegans with autoregressive hidden Markov models
E. Kelly Buchanan, Akiva Lipshitz, Scott Linderman, Liam Paninski. *NeurIPS, 2017, Workshop on Worm's Neural Information Processing*. **Selected for a spotlight presentation**.

Selected Honors and Awards

- Open Benchmarks Grant, Snorkel AI 2026
- Entropic Award for Most Surprising Negative Result, ICBINB Workshop at NeurIPS 2022
- Innovation Award, Google X 2022
- Ph.D. Engineering and Science Fellowship, GEM Consortium 2019
- Machine Learning Summer School Scholarship, Imperial College London 2019
- Deep Learning and RL Summer School Scholarship, Amii / Vector Institute 2018–2019
- Wallace S. Strobel Fellowship, University of Kansas 2014–2016
- David D. and Mildred H. Robb Award, University of Kansas School of Engineering 2015
- Full Tuition Award for Academic Excellence, Univ. of Kansas and Fulbright Peru 2009–2014

Invited Talks

Catching Benchmark Rot: A Systematic Audit of Terminal-Bench
Z.AI (April 2026)· Cursor (May 2026)· Mirendil (June 2026).

Developing Reliable Frameworks for Longitudinal Evaluation of AI Agents
Together AI (Oct 2025).

Pathologies of Predictive Diversity in Deep Ensembles
NYU Center for Data Science (Oct 2023).

Building Reliable Tools for Neuroscience Research
Stanford University (Sep 2023) · Gatsby Tri-Center Meeting (Jun 2023) · Caltech (Mar 2023) · MBZUAI (Jan 2023).

Deep Ensembles Work, but Are They Necessary?
Amazon Science (Mar 2023) · ETH Zurich (Feb 2023) · NYU Abu Dhabi (Jan 2023).

Plex: Last-Layer Changes in CLIP Models for Semantic Segmentation
Columbia University (Oct 2022) · Google AI (Sep 2022).

Algoritmos para Acelerar el Descubrimiento Científico en la Neurociencia
TECHSUYO (Oct 2021).

Encoding Expert Knowledge for Improved Behavioral Video Analysis
Araya Research (Jan 2021) · Columbia U19 Motor Control (Nov 2020) · Neuromatch 3.0 (Nov 2020).

Denoising and Demixing Neural Activity in Functional Imaging Data
UCL Cortex Lab (Jul 2019) · IARPA MICrONS Meeting (Mar 2018).

Challenges in Online Spike Detection and Clustering during Electrical Microstimulation
Italian Institute of Technology (Jan 2016) · KU Medical Center (Dec 2015).

Abstracts

[1] **Faster Neural to Speech Decoding with Deep State Space Models**.
E. Kelly Buchanan, Hyun Dong Lee, David Zoltowski, Scott W. Linderman.
Computational and Systems Neuroscience. Lisbon, Portugal. 2026.

[2] **Constrained matrix factorization methods for denoising and demixing voltage imaging data**.
E. Kelly Buchanan, Johannes Friedrich, Ian Kinsella, Patrick Stinson, Pengcheng Zhou, Felipe Gerhard, John Ferrante, Graham Dempsey, Liam Paninski.
Computational and Systems Neuroscience. Denver, CA. 2018.

[3] **Validity of extracellular recordings under spike-triggered stimulation**.
E. Kelly Buchanan, David J. Guggenmos, Gustaf M. Van Acker, Randolph Nudo.
Society for Neuroscience Annual Meeting. San Diego, CA. November 2016.

[4] **Efficacy of an automated spike detection algorithm for processing in vivo multichannel recordings**.
E. Kelly Buchanan, David J. Guggenmos, Alverto Averna, Caleb Dunham, Gustaf Van Acker, Randolph J. Nudo, Michaela Chiappalone.
Society for Neuroscience Annual Meeting. Chicago, IL. October 2015.

Teaching

- Guest Lab, Foundations of Graphical Models, Columbia University — Fall 2020
- Supplemental Instructor, Physics I and Introduction to Digital Logic Design, Univ. of Kansas — Fall 2010–2013

Professional Activities & Service

- Co-organizer of I Can't Believe It Is Not Better Workshop at NeurIPS — 2023–2024
- Co-organizer of The Symbiosis of Deep Learning and Differential Equations at NeurIPS — 2021–2022

- Reviewer for JMLR (2023), ICML (2021–2024), NeurIPS (2020–), Cosyne (2025–)
- Senior Program Chair at Women in Machine Learning 2023–2024
- Co-founded the Culture and Belonging Committee at Columbia University Theory Center 2019–2022
- Co-founded Girls Who Code at Zuckerman Institute 2019–2020
- Coding tutor for middle school children at CoderDojo KC 2015–2016