Kempner INSTITUTE

For the Study of Natural & Artificial Intelligence at Harvard University

Research Roundup

March 2026

Discover our latest insights in natural and artificial intelligence research

- Associate Faculty Venkatesh Murthy and his collaborators published a paper in *PNAS* showing that "**Convergent Motifs of Early Olfactory Processing Are Recapitulated by Layer-Wise Efficient Coding**."
- Research Fellow Binxu Wang and Affiliate Faculty Carlos Ponce published a **study** in *Nature Neuroscience* that "showed that neurons in the primate ventral stream can successfully guide image optimizations (evolutions) in two generative image spaces."
- A team including Institute Investigator Yilun Du and incoming Institute Investigator Michael Albergo introduced "**Graph Energy Matching: Transport-Aligned Energy-Based Modeling for Graph Generation.**"

Featured Formula

$$r(\hat{y}, c) = \underbrace{2\phi_c(\hat{y})^\top \phi_c(y)}_{\text{alignment term}} - \underbrace{2\phi_c(\hat{y})^\top \mathbb{E}_{\tilde{y}\sim p_\theta(\cdot|c)}\left[\phi_c(\tilde{y})\right]}_{\text{diversity term}}$$

Formula decomposing reward into an alignment term and a diversity term, from the preprint: "**Matching Features, Not Tokens: Energy-Based Fine-Tuning of Language Models.**"

From the Deeper Learning **blog post** about the work: "The alignment term rewards completions whose features are close to the ground truth — pushing the model toward semantically faithful outputs. The diversity term penalizes completions that are too similar to each other — preventing the model from collapsing onto a single mode."

Featured Figure

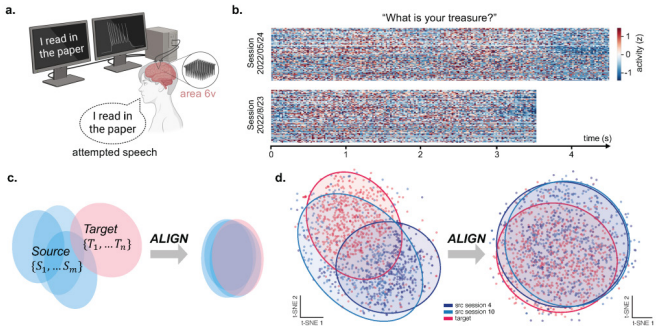

Diagram from the preprint "**ALIGN: Adversarial Learning for Generalizable Speech Neuroprosthesis**," presenting an overview of how the ALIGN framework addresses session-to-session variability in neural recordings during attempted speech.

Featured Artifacts

Matching Features, Not Tokens: Energy-Based Fine-Tuning of Language Models.
Samy Jelassi, Mujin Kwun, Rosie Zhao, Yuanzhi Li, Nicolo Fusi, Yilun Du, Sham M. Kakade, and Carles Domingo-Enrich.

Get the Code ▶

Pitfalls in Evaluating Interpretability Agents.
Tal Haklay, Nikhil Prakash, Sana Pandey, Antonio Torralba, Aaron Mueller, Jacob Andreas, Tamar Rott Shaham, and Yonatan Belinkov.

Get the Code ▶

Jailbreak Scaling Laws for Large Language Models: Polynomial-Exponential Crossover.
Indranil Halder, Annesya Banerjee, and Cengiz Pehlevan.

Get the Code ▶

AI Innovation

See It to Place It: Evolving Macro Placements With Vision-Language Models.
Ikechukwu Uchendu, Swati Goel, Karly Hou, Ebrahim Songhori, Kuang-Huei Lee, Joe Wenjie Jiang, Vijay Janapa Reddi, and Vincent Zhuang. arXiv:2603.28733v1 (2026).

Graph Energy Matching: Transport-Aligned Energy-Based Modeling for Graph Generation.
Michal Balcerak, Suprosana Shit, Chinmay Prabhakar, Sebastian Kaltenbach, Michael S. Albergo, Yilun Du, and Bjoern Menze. arXiv:2603.23398v1 (2026).

Matching Features, Not Tokens: Energy-Based Fine-Tuning of Language Models.
Samy Jelassi, Mujin Kwun, Rosie Zhao, Yuanzhi Li, Nicolo Fusi, Yilun Du, Sham M. Kakade, and Carles Domingo-Enrich. arXiv:2603.12248v2 (2026).

Learning Adaptive LLM Decoding.
Chloe H. Su, Zhe Ye, Samuel Tenka, Aidan Yang, Soonho Kong, and Udaya Ghai. arXiv:2603.09065v2 (2026).

A Monte Carlo Estimator of Flow Fields for Sampling and Noise Problems.
Michael S. Albergo and Gurtej Kanwar. arXiv:2603.00252v1 (2026).

Interpretability and AI Theory

Pitfalls in Evaluating Interpretability Agents.
Tal Haklay, Nikhil Prakash, Sana Pandey, Antonio Torralba, Aaron Mueller, Jacob Andreas, Tamar Rott Shaham, and Yonatan Belinkov. arXiv:2603.20101v1 (2026).

Jailbreak Scaling Laws for Large Language Models: Polynomial-Exponential Crossover.
Indranil Halder, Annesya Banerjee, and Cengiz Pehlevan. arXiv:2603.11331v1 (2026).

Learning to Rank the Initial Branching Order of SAT Solvers.
Arvid Eriksson, Gabriel Poesia, Roman Bresson, Karl Henrik Johansson, and David Broman. arXiv:2603.07176v1 (2026).

Why Depth Matters in Parallelizable Sequence Models: A Lie Algebraic View.
Gyuryang Heo, Timothy Ngotiaoco, Kazuki Irie, Samuel J. Gershman, and Bernardo Sabatini. arXiv:603.05573v1 (2026).

Applications of AI

A Deep Learning Framework for Predicting Teprotumumab Treatment Response in Thyroid Eye Disease.
Saul Langarica, Nahyoung Grace Lee, Adham M. Alkhadrawi, Young-Tak Kim, Sierra K. Ha, Synho Do, and Lisa Y. Lin. Ophthalmology Science 6(4) :101098 (2026).

Applications of AI

Graph Neural Network Modeling of Spatial Tumor-Immune Interactions Identifies Prognostic Cellular Niches in Non-Small Cell Lung Cancer.
Katharina V. Hoebel, James R. Lindsay, Jennifer Altreuter, Joao V. Alessi, Jason L. Weirather, Ian Dryg, Anita Giobbie-Hurder, Zhirou Li, Kun-Hsing Yu, Mark M. Awad, Scott J. Rodig, and William Lotter. NPJ Precision Oncology (2026).

Neuroscience & NeuroAI

Convergent Motifs of Early Olfactory Processing Are Recapitulated by Layer-Wise Efficient Coding.
Juan Carlos Fernández del Castillo, Farhad Pashakhanloo, Venkatesh N. Murthy, and Jacob A. Zavatone-Veth. Proceedings of the National Academy of Sciences of the United States of America 123(13): e2524661123 (2026).

Phasic Dopamine Drives Conditioned Responding Beyond its Role in Learning.
Jay A. Hennig, Mark Burrell, Naoshige A. Uchida, and Samuel J. Gershman. bioRxiv:2026.03.25.714259 (2026).

ALIGN: Adversarial Learning for Generalizable Speech Neuroprosthesis.
Zhanqi Zhang, Shun Li, Bernardo L. Sabatini, Mikio Aoi, and Gal Mishne. arXiv:2603.18299v1 (2026).

Linear Readout of Neural Manifolds with Continuous Variables.
Will Slatton, Chi-Ning Chou, and SueYeon Chung. arXiv:2603.10956v1 (2026).

Neuronal Tuning Aligns Dynamically With Object and Texture Manifolds Across the Visual Hierarchy.
Binxu Wang and Carlos R. Ponce. Nature Neuroscience (2026).

Structure, Disorder, and Dynamics in Task-Trained Recurrent Neural Circuits.
David G. Clark, Blake Bordelon, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. bioRxiv:2026.03.02.708943 (2026).

Data-Derived Agents Reveal Dynamical Reservoirs in Mouse Cortex for Adaptive Behavior.
Siyan Zhou, Ryan P. Badman, Charlotte Arlt, Kanaka Rajan, and Christopher D. Harvey. bioRxiv:2026.03.03.709365v1 (2026).

Note: *This is a partial list of articles and preprints published by Kempner-affiliated researchers in the last month. Papers are listed by topic and publication/upload date, with the most recent first.*

View this report and other Kempner Research-Round-ups:
kempnerinstitute.harvard.edu/kempner-community/research-roundup/