Kempner INSTITUTE

For the Study of Natural & Artificial Intelligence at Harvard University

Research Roundup

June 2026

Discover our latest insights in natural and artificial intelligence research

- A team including Kempner Graduate Fellow Chloe Huangyuan Su, Kempner Co-Director Sham Kakade and Institute Investigator Yilun Du published a **preprint** that examines an "agent economy in which agents compete via auctions for the right to act, exchange payments, and accumulate wealth from environmental rewards."

- Kempner Research Fellow William Dorrell presents a **new theory** for understanding Sparse Autoencoders (SAEs), using it to explain a range of observed SAE behaviors: "hierarchical splitting & absorption, the structure of residuals, and dense antipodal features."

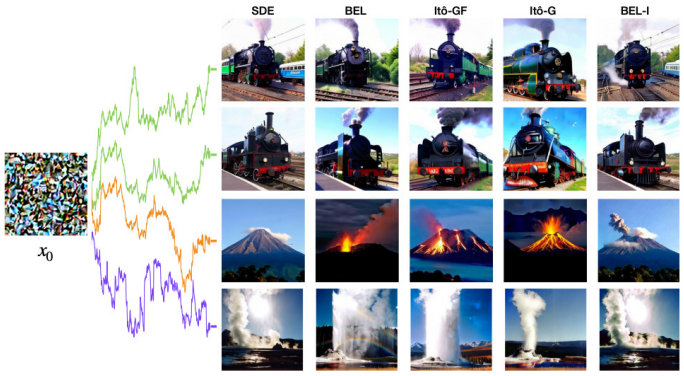
- A team including Research Fellow Richard Hakim and associate faculty members Venkatesh Murthy and Demba Ba introduces **NEURRATOR**, "a framework that decodes spiking activity into free-form natural-language narration of the viewed scene at single-neuron resolution."

Featured Formula

$$\mathcal{E}_t = \sum_{p=0}^{t}\sum_{q=0}^{t}\binom{t}{p}\binom{t}{q}\left(-\frac{1}{\alpha}\right)^{p+q}\frac{1}{p!\,q!}\frac{\partial^p}{\partial u^p}\frac{\partial^q}{\partial v^q}F(u,v)\Big|_{u=v=0}$$

Generalization error at chain of thought depth t, from the preprint "**An Asymptotic Theory of Chain-of-Thought in In-Context Learning.**"

Featured Figure



Examples of inference-time steering of an image generation model, from the preprint **Itô Maps for Any-Step SDEs.** The authors introduce the Itô map, "which can be thought of as path-wise stochastic flow maps whose stochasticity is driven by the Brownian motion of the underlying SDE. "

Featured Artifacts

Can Neurons Speak? Semantic Narration of Vision at Single-Cell Resolution.
Arnau Marin-Llobet, Richard Hakim, Sara Matias, Venkatesh N. Murthy, Na Li, and Demba Ba (2026).

Get the Code ▶

Economy of Minds: Emerging Multi-Agent Intelligence with Economic Interactions.
Zhenting Qi, Huangyuan Su, Ao Qu, Chenyu Wang, Yu Yao, Han Zheng, Kushal Chattopadhyay, Guowei Xu, Zihan Wang, Weirui Ye, Vijay Janapa Reddi, Ju Li, Paul Pu Liang, Himabindu Lakkaraju, Sham Kakade, and Yilun Du. (2026).

Get the Code ▶

kempnerinstitute.harvard.edu

AI Innovation

Temporal Backtracking Search for Test-time Generative Video Reasoning.
Sejoon Jun, Zheng Ding, Huangyuan Su, Weirui Ye, and Yilun Du. arXiv:2606.13861v1 (2026).

Itô Maps for Any-Step SDEs.
Zhengkai Pan, Peter Potaptchik, Wenxi Yao, Michael S. Albergo, and Jakiw Pidstrigach. arXiv:2606.11156v1 (2026).

Low-Frequency Shortcuts in Texture-Driven Visual Learning.
Utku Şirin, Cathy Hou, David Alvarez-Melis, and Stratos Idreos. arXiv:2606.03493v1 (2026).

Economy of Minds: Emerging Multi-Agent Intelligence with Economic Interactions.
Zhenting Qi, Huangyuan Su, Ao Qu, Chenyu Wang, Yu Yao, Han Zheng, Kushal Chattopadhyay, Guowei Xu, Zihan Wang, Weirui Ye, Vijay Janapa Reddi, Ju Li, Paul Pu Liang, Himabindu Lakkaraju, Sham Kakade, and Yilun Du. arXiv:2606.02859v1 (2026).

Interpretability and AI Theory

Compute Efficiency and Serial Runtime Tradeoffs for Stochastic Momentum Methods.
Depen Morwani, Alexandru Meterez, Pranav Nair, and Sham Kakade. arXiv:2606.19179v1 (2026).

Adversarial Concept Search: Predicting Compositional Errors From Feature Geometry.
Jennifer Meng Lu, Ruochen Zhang, Isabelle Lee, David Alvarez-Melis, Ellie Pavlick, and Naomi Saphra. arXiv:2606.13934v1 (2026).

A Unifying View of Attention Sinks: Two Algorithms, Two Solutions.
Lukas Fesser, Mozes Jacobs, Thomas Fel, Andy Keller, and Sham Kakade. arXiv:2606.08105v1 (2026).

Where the Score Lives: A Wavelet View of Diffusion.
Emma Finn, Binxu Wang, T. Anderson Keller, and Demba E. Ba. arXiv:2606.08309v1 (2026).

Position: Don't Just "Fix it in Post": A Science of AI Must Study Training Dynamics.
Stella Biderman, Mohammad Aflah Khan, Niloofar Mireshghallah, Catherine Arnett, Fazl Barez, and Naomi Saphra. arXiv:2606.06533v1 (2026).

An Asymptotic Theory of Chain-of-Thought in In-Context Learning.
Kaito Takanami and Cengiz Pehlevan. arXiv:2606.03217v1 (2026).

Interpretability and AI Theory

How Optimality Structures Sparse Dictionaries: A Theory for Understanding SAE Representations.
William Dorrell. arXiv:2606.02385v1 (2026).

Task-Induced Representational Invariances Depend on Learning Objective in Deep RL.
Manu Srinath Halvagal, Sebastian Lee, and SueYeon Chung. arXiv:2606.01868v1 (2026).

Applications of AI

How Post-Training Shapes Biological Reasoning Models.
Lukas Fesser, Hanlin Zhang, Michelle M. Li, Eric Wang, Bryan Perozzi, Shekoofeh Azizi, Sham M. Kakade, and Marinka Zitnik. arXiv:2606.16517v1 (2026).

Stable Menus of Public Goods: AI-Enabled Progress.
Sara Fish. arXiv:2606.16989v1 (2026).

Embeddings of Clinical Codes Enable Knowledge-Grounded AI in Medicine.
Ruth Johnson, Uri Gottlieb, Galit Shaham, Lihi Eisen, Jacob Waxman, Stav Devons-Sberro, Curtis R. Ginder, Peter Hong, Raheel Sayeed, Xiaorui Su, Ben Y. Reis, Ran D. Balicer, Noa Dagan, and Marinka Zitnik. NPJ Digital Medicine (2026).

Neuroscience & NeuroAI

Can Neurons Speak? Semantic Narration of Vision at Single-Cell Resolution.
Arnau Marin-Llobet, Richard Hakim, Sara Matias, Venkatesh N. Murthy, Na Li, and Demba Ba. arXiv:2606.18667v1 (2026).

Implications of Hierarchical Markov Models of Behavior: On Irreversibility, Predictability, and Dimensionality.
John J. Vastola and Kanaka Rajan. arXiv:2606.14692v1 (2026).

Interpreting Brain Responses to Language with Sparse Features from Language Models.
Michael A. Lepori, Kendrick Kay, and Greta Tuckute. arXiv:2606.06857v1 (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/