

Wannan (Winnie) Yang

📍 New York and Seattle, USA ✉️ winnieyangwn96@gmail.com 🌐 [personal website](#) [in linkedin](#) [github](#)

EDUCATION

Ph.D. candidate in Computational Neuroscience. New York University, Buzsáki Lab 🔗	Graduating in 2025.9
Visiting Student. MIT, Tye Lab 🔗	2018.6 – 2019.6
B.S. in Computational Neuroscience. University of Edinburgh. <i>GPA: 4.0 (USA equivalent)</i>	2014.9 – 2018.5

RESEARCH EXPERIENCE

LLM Interpretability and Alignment (ICLR 🔗)	2024.3 – present
--	-------------------------

Collaborator: [Chen Sun](#), Google DeepMind

- Designed and conducted experiments to study two safety-related problems in large language models (LLMs): *deception* and *jailbreaks*.
- Implemented a battery of interpretability tools including contrastive activation steering, activation patching and sparse auto-encoders (SAEs) to understand and control LLMs.
- Built a pipeline ([github](#) [🔗](#)) to evaluate, analyze and steer 25+ large language models from different model families (Gemma, Llama, Pythia, Qwen and Yi) of different sizes (form 1.5 billion to 70 billion parameters).
- Published a series of technical [blog posts](#) [🔗](#) to share the research findings.

Memory Representation and Consolidation (Science 🔗, Nature 🔗, NeurIPS 🔗)	2020.9 – present
---	-------------------------

Mentor: [György Buzsáki](#), NYU

- Led a project to study a key mechanism for selective memory consolidation in the brain. This novel discovery has led to a publication in *Science* (leading author).
- Developed a novel latent-space based decoding method and applied various ML tools (including Bayesian decoding) to decode the content of memory reactivations (‘replays’) from neural population activity during learning and sleep.
- Created a pipeline for decoding large-scale (50TB) electrophysiology data. The pipeline can be applied to a diverse range of tasks. Implemented variants of the method to different datasets and projects, which enabled further key publications, including a collaboration project recently accepted at *Nature* (in press) and a first author paper at *NeurIPS* Symmetry and Geometry in Neural Representations Workshop.
- Open-sourced [demo codes](#) [🔗](#) and [tutorials](#) [🔗](#). The neural data processing and decoding pipeline has been widely used by lab members and colleges from other research labs.

Brain-inspired Deep Reinforcement Learning (NeurIPS 🔗)	2021.3 – 2023.9
---	------------------------

Collaborator: [Chen Sun](#), Google DeepMind

- Co-developed a brain-inspired (memory consolidation and reflection) framework to build a novel deep RL algorithm.
- The resulting simple and scalable algorithm greatly improved long-term credit assignments in a diverse set of RL tasks (including grid-world, Montezuma’s Revenge and other Atari games).

PUBLICATIONS

-
- [W. Yang](#), [C. Sun](#), [G. Buzsáki](#). (2024). **Interpretability for Safe AI: Jailbreak as a case study.** *In preparation*.
 - [W. Yang](#), [G. Buzsáki](#). (2024). **Interpretability of LLMs Deception: Universal Motif.** *ICLR* [🔗](#) (under review).
 - [W. Yang](#), [C. Sun](#), [R. Huszár](#), [T. Hainmueller](#), [K. Kiselev](#), [G. Buzsáki](#). (2024). **Selection of experience for memory by hippocampal sharp wave ripple.** *Science* 383, 1478-1483. [🔗](#)
 - [I. Zutshi](#), [A. Apostolelli](#), [W. Yang](#), [Z. Zheng](#), [T. Dohi](#), [E. Balzani](#), [A. H. Williams](#), [C. Savin](#), [G. Buzsáki](#). (2024). **Hippocampal neuronal activity is aligned with action plans.** *Nature* (in press) [🔗](#).
 - [C. Sun](#), [W. Yang](#), [T. Jiralerspong](#), [D. Malenfant](#), [B. Alsbury-Nealy](#), [Y. Bengio](#), [B. Richards](#). (2023). **Contrastive Retrospection: honing in on critical steps for rapid learning and generalization in RL.** *NeurIPS*. [🔗](#)
 - [W. Yang](#), [C. Sun](#), [R. Huszár](#), [G. Buzsáki](#). (2023). **Changes in the geometry of hippocampal representations across brain states.** Symmetry and Geometry in Neural Representations Workshop *NeurIPS*. [🔗](#)
 - [E. Y. Kimchi](#), [A. Burgos-Robles](#), [G. A. Matthews](#), [T. Chakoma](#), [M. Patarino](#), [J. Weddington](#), [C. A. Siciliano](#), [W. Yang](#), [S. Foutch](#), [R. Simons](#), [M. Fong](#), [M. Jing](#), [Y. Li](#), [D. B. Polley](#), [Kay M. Tye](#). (2023). **Reward contingency gates selective cholinergic suppression of amygdala neurons.** *eLife* [🔗](#)
 - [S. Tennant](#), [I. Hawes](#), [H. Clark](#), [W. Tam](#), [J. Hua](#), [W. Yang](#), [K. Gerlei](#), [E. Wood](#), [M. Nolan](#). (2022). **Analogue representation of a spatial memory by ramp-like neural activity in retrohippocampal cortex.** *Current Biology* [🔗](#)
 - [C. Sun](#), [W. Yang](#), [J. Martin](#), [S. Tonegawa](#). (2020). **Hippocampal neurons represent events as transferable units of experience.** *Nature Neuroscience* [🔗](#).

SKILLS

ML: Pytorch, scikit-learn








LLM Agent: LangChain, AutoGen

LLM Interpretability: transformer-lens, Huggingface Transformers, Contrastive Activation Steering, Activation Patching, SAE Steering

Programming: Python, MATLAB, HTML, LaTeX

Computational Neuroscience: Large-scale High-dimensional Data Analysis, Linear and Nonlinear Dimensionality Reduction, Time Series Data Analysis, Neural Data Decoding

COURSES

Large Language Model Agents Instructor: Dawn Song 	Ongoing
Deep Learning Instructor: Yann LeCun 	NYU. Grade: A
Computational Cognitive Modeling Instructor: Brenden Lake 	NYU. Grade: A
Reinforcement Learning Instructor: David Silver 	UCL.
Neural Circuits and Computational Modeling Instructor: Xiaojing Wang 	NYU. Grade: A
Neural Networks and Deep Learning Instructor: Andrew Ng 	deeplearning.ai
Applied Machine Learning Instructor: Oisín Mac Aodha 	UoE. Grade: A