Wannan(Winnie) **Yang**

Computational Neuroscience · LLM Interpretability · LLM Agent

winnieyangwn96@gmail.com | https://winnieyangwannan.github.io/ | winnieyangwannan | @winnieyangwn | Wannan(Winnie) Yang

Education

New York University
New York, USA

PhD in Computational Neuroscience
Graduating Sep. 2025

- PhD candidate at the Buzsaki Lab.

University of Edinburgh
Edinburgh, UK

B.S. in Computational Neuroscience
Sep. 2014 - Jun. 2019

- First Class with honors (highest class, US equivalent GPA: 4.0).
- Thesis Project conducted in the Nolan Lab.

Massachusetts Institute of Technology
Boston, US

Visiting student
Jun. 2017 - Jun. 2018

- Visiting student (partially supported by the Principal's Go Abroad Award).
- Research assistent in the Tye Lab.

Research Experience

Large Language Model Interpretability (NeurIPS-1, NeurIPS-2)
NYU, U.S.A

Collaborating with Chen Sun from Google DeepMind
May 2024 - Present

- Designed and conducted evaluation and interpretability experiments to study two significant safety-related problems in large language models (LLMs): *deception* (paper link) and *jailbreaks* (paper link). This work led to two first arthur papers, submitted to the Safe Generative AI workshop at ***NeurIPS*** .
- Implemented a battery of tools including contrastive activation steering, activation patching and sparse auto-encoders (SAEs) to analyze and control LLMs.
- Built a **pipeline** to evaluate, analyze and steer 25+ open-source large language models of different model families (Gemma, Llama, Pythia, Qwen and Yi) of different sizes ranging form 1.5 billion to 70 billion parameters.
- Published a series of blog posts to share the research findings.

Memory Consolidation and Neural Representation in the Brain (Science, Nature, NeurIPS)
NYU, U.S.A.

Buzsaki Lab
Sep. 2020 - Present

- Led a project to study the mechanism for selective memory consolidation in the brain. This 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 reactivation (replays) from neural population activity during learning and sleep.
- Created a pipeline for decoding large-scale (>50 TB) 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 publications include a collaboration project recently accepted in ***Nature*** (in press) and a first author paper at ***NeurIPS*** Workshop on Symmetry and Geometry in Neural Representations
- Open-sourced demo codes and tutorials. The neural data processing and decoding pipeline has been been widely used by lab members and colleges from other research labs.

Skills

ML	Pytorch, scikit-learn
LLM	LangChain, transformer-lens, Hugging Face Transformers
Programming	Python, MATLAB, HTML, LaTeX
Research	Large-scale high-Dimensional Data Analysis, Time Series Data Analysis, Neural Data Decoding

Publications

Wannan Yang, Chen Sun, György Buzsáki.
Model Interpretability and Model Steering for safe AI: A Case Study on jailbreaks. NeurIPS Safe Generative AI Workshop (2024).
NeurIPS link

Wannan Yang, Zhuonan (Jojo) Yang, Chen Sun, György Buzsáki.
Model Interpretability and Model Steering for safe AI: A Case Study on deception. NeurIPS Safe Generative AI Workshop (2024).
NeurIPS link

Wannan Yang, Chen Sun, Roman Huszár, Thomas Hainmueller, Kirill Kiselev, György Buzsáki.
Selection of experience for memory by hippocampal sharp wave ripple. Science 383, 1478-1483 (2024).
project website Quanta article

Ipshita Zutshi, Athina Apostolelli, Wannan Yang, Zheyang Zheng, Tora Dohi, Edoardo Balzani, Alex H Williams, Cristina Savin, György Buzsáki.
Hippocampal neuronal activity is aligned with action plans. Nature (in press) (2024).
preprint link

Chen Sun, Wannan Yang, Thomas Jiralerspong, Dane Malenfant, Benjamin Alsbury- Nealy, Yoshua Bengio, Blake Richards.
Contrastive Retrospection: honing in on critical steps for rapid learning and generalization in RL. NeurIPS (2023).
NeurIPS link

Wannan Yang, Chen Sun, Roman Huszár, György Buzsáki.
Changes in the geometry of hippocampal representations across brain states. NeurIPS Workshop on Symmetry and Geometry in Neural Representations (2023).
NeurIPS link

Eyal Y. Kimchi , Anthony Burgos-Robles, Gillian A. Matthews, Tatenda Chakoma, Makenzie Patarino, Javier Weddington, Cody A. Siciliano, Wannan Yang, Shaun Foutch, Renee Simons, Ming-fai Fong, Miao Jing, Yulong Li, Daniel B. Polley, Kay M. Tye.
Reward contingency gates selective cholinergic suppression of amygdala neurons. eLife (2023).
pdf link

Sarah A. Tennant , Ian Hawes, Harry Clark, Wing Kin Tam, Junji Hua, Wannan Yang, Klara Gerlei, Emma R. Wood, Matthew F. Nolan.
Analogue representation of a spatial memory by ramp-like neural activity in retrohippocampal cortex. Current Biology (2022).
pdf link

Chen Sun, Wannan Yang, Jared Martin, Susumu Tonegawa.
Hippocampal neurons represent events as transferable units of experience. Nature Neuroscience (2020).

Awards

Amgen Scholarship

2018

- Received funding for conducting independent research, acceptance rate around 7%.
- Hosted by the Gogolla Lab at Max Planck Institute in Germany.

Principal's Go Abroad Award

2017

- Received the funding to do research as visiting student at the Tye lab at MIT.

Selected Courses

Large Language Model Agents *Ongoing*
Instructer: Dawn Song *2024*

Deep Learning *NYU. Grade: A*
Instructer: Yann LeCun. *2022*

Computational Cognitive Modeling *NYU. Grade: A*
Instructer: Brenden Lake *2022*

Reinforcement Learning *Online*
Instructor: David Silver *2020*

Neural Circuits and Computational Modeling *NYU, Grade: A*
Instructor: Xiaojing Wang *2019*

Neural Networks and Deep Learning *Online*
Instructor: Andrew Ng *2018*

Applied Machine Learning *University of Edinburgh. Grade: A*
Instructor: Oisin Mac Aodha *2018*