Wannan (Winnie) Yang

San Francisco, USA winnieyangwn96@gmail.com personal website Google Scholar github

My work spans from biological brains (my works were published in Science and Nature) to autonomous LLM research agents (current research at Meta). I've spent years understanding biological learning machines and currently building artificial ones. The brain taught me how intelligence works. AI taught me how to build it. Now I want to close this circle—using insights from biological intelligence to build better AI, then wielding that AI to decode biology itself.

EDUCATION

Ph.D. candidate. New York University — Graduating in Spring 2026
Visiting Student. MIT — 2018.6 – 2019.6
B.S. in Computational Neuroscience. University of Edinburgh. *GPA: 4.0 (USA equivalent)* — 2014.9 – 2018.6

RESEARCH EXPERIENCE

AI Research Agent (AIRA) — Starting 2025.11

Student Researcher at Meta FAIR, Manager: Yoram Bachrach

- Will join the AIRA Team at Meta FAIR to build toward a fully automated AI research scientist. Focusing on building LLM research agents and hill-climb on research capabilities (MLE-Bench etc.)

Self-improving and domain-general LLM Coding Agent — 2025.6 – present

Student Researcher at Meta, Collaborator: Jenny Zhang

- Developing an *Evolutionary Algorithm-based* meta learning Agent that autonomously generates **domain-general** task agents across various domains, achieving strong performance across diverse challenges including NetHack (game-play domain), International Mathematical Olympiad (IMO) problems (math domain), and automated paper review (non-verifiable domain).
- Developing an **LLM-based code agent** that autonomously designs reward functions for solving embodied AI tasks with RL, including locomotion and robotic manipulation, using the Genesis simulator.

LLM Alignment and Post Training (MI NeurIPS) — 2025.3 – 2025.08

Research Scientist Internship at Meta

- Led a research project to develope a novel post-training algorithm to mitigate hallucination in LLMs.
- The new algorithm delivers ∼ **30x higher compute efficiency (FLOPs per token)** and requires ∼**20x less training data** comparing to competitive baselines like SFT and DPO.
- Developed a modality-agnostic training pipeline, effectively mitigating hallucination in both text-only and multimodal models.
- Proposed *the first* ever steering-based training framework with general applicability to **both dense and Mixture-of-Experts (MoE) models**.

LLM Alignment and Interpretability (preprint) — 2024.3 – 2025.3

Collaborator: Chen Sun, Google DeepMind

- Designed and conducted experiments to study *deception* in LLMs.
- Built a open-source pipeline (github) that allow easy hypothesis testing and fast experiments to evaluate, analyze and steer 25+ large language models from different model families (Gemma, Llama, Qwen and Yi) of different sizes (from 1.5 billion to 405 billion parameters).

Memory Representation and Consolidation (Science, Nature, NeurIPS) — 2022.9 – 2024.3

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').
- Created a pipeline for decoding large-scale ($50TB$) electrophysiology data.
- Implemented variants of the method to different datasets and projects, which enabled further key publications, including a collaboration project published in *Nature* and a first author paper at *NeurIPS* Workshop.

Brain-inspired Deep Reinforcement Learning (NeurIPS) — 2021.9 – 2022.9

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, X. Qiu, L. Yu, Y. Zhang, O. A. Yang, N. Kokhlikyan, N. Cancedda, D. Garcia-Olano. (2025). **Hallucination Reduction with CASAL: Contrastive Activation Steering for Amortized Learning** *MI NeurIPS*.
- W. Yang, Z. Yang, C. Sun, G. Buzsáki. (2025). **How Large Language Models Lie: Rotation of the Truth Direction as a Universal Motif.** *preprint*.
- I. Zutshi, A. Apostolelli, W. Yang, Z. Zheng, T. Dohi, E. Balzani, A. H. Williams, C. Savin, G. Buzsáki. (2025). **Hippocampal neuronal activity is aligned with action plans.** *Nature*.
- W. Yang, C. Sun, G. Buzsáki. (2024). **Interpretability for Safe AI: LLM Lying as a case study**. *NeurIPS* (SafeGenAi Workshop).
- W. Yang, C. Sun, G. Buzsáki. (2024). **Interpretability for Safe AI: Jailbreak as a case study**. *In preparation.*
- 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.
- 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

LLM Agent: Coding Agent; Large Scale Multi-agent, Multi-turn, Multi-task Agentic Systems

LLM Training: SFT; online and offline RL training for LLMs (DPO, PPO, GRPO, etc.); distributed training across GPUs and on accelerators; RL for robotics control tasks (with Genesis simulation platform)

LLM Interpretability and Alignment: transformer-lens, Contrastive Activation Steering, Activation Patching, SAE Steering

Programming: Python, MATLAB, HTML, LaTeX

Statistical Data Analysis: Large-scale High-dimensional Data Analysis, Signal processing, Linear and Nonlinear Dimensionality Reduction, Time Series Data Analysis on biological data, Neural Data Decoding, Multimodal Data Analysis