Jingxuan Fan (She/Her/hers)

Boston, MA·jfan@g.harvard.edu·+1(909)344-4142
linkedin.com/in/jingxuan-fan/ · github.com/jingxuanf0214

EDUCATION

Harvard University **Cambridge, MA**
Ph.D. Candidate, Program in Neuroscience Expected Jan. 2026
Dissertation: Reinforcement Learning with Dense Intrinsic Rewards for Complex Environment Navigation
M.S. in Applied Math March 2024
Relevant Coursework: Reinforcement Learning, Neural Computation, Advanced Topics in Data Science, Physical Mathematics, TinyML and Efficient Deep Learning, Mathematical & Engineering Principles for Training Foundation Models

Massachusetts Institute of Technology **Cambridge, MA**
B.S. in Brain and Cognitive Science May 2020
Honors & Awards: Hans Lukas Teuber Award for Outstanding Academics, Walle J.H. Nauta Award for Outstanding Research

SKILL & INTERESTS

Programming skills: Python (PyTorch, Tensorflow, scikit-learn, Pandas, SciPy/NumPy), Matlab, SQL, Linux, Git
Modeling skills: LLM post-training, reinforcement learning, agent, text-to-image diffusion, mechanistic interpretability

SELECTED RESEARCH EXPERIENCE

Harvard University ***Cambridge, MA***
PhD Researcher, Dept. of Neurobiology *Expected Jan. 2026*

- **Developed an automatic framework to generate a simple synthetic dataset** and **benchmarked** a novel and important LLM behavior – information bias along user-assistant axis – across 26 open-source models (base, instruction-tuned, reasoning) and 26 closed-source models (non-reasoning, reasoning); benchmarking result demonstrates how different post-training methods result in different user-assistant bias; **performed RLHF or reasoning trace SFT** on Llama and Qwen family base models and comprehensively demonstrated user-assistant bias evolving over training stages; **finetuned on the synthetic dataset** can showed bidirectionally change in models' user-assistant bias on real-life conversations (submission to **ICLR 2026**)
- **Pretrained text-to-image diffusion model** on carefully designed text-image corpora and conducted controlled attention localization, ablation and circuit discovery to understand the attention mechanisms for generating different object properties – color, shape and spatial relationship; **Discovered a general solution** for generating correct object spatial relationships **and demonstrated** how convergence on this solution vary with text encoding and parameter size (**NEMI workshop**, submission to **ICLR 2026**)
- **Developed a novel RL post-training pipeline** – reward model training, benchmarking and policy model training – to improve math domain specific question-answer (QA) performance with only web sourced math text (finemath), bypassing the need for resource-consuming dataset construction

Master's Researcher, Dept. of Applied Mathematics *March 2024*

- Developed an **entropy-penalized composition method** for multi-attribute reward models and demonstrated **improved results on reward model benchmarks** (submission to **AAAI 2026**)
- **Developed** a framework to **generate large-scale synthetic rule pool** and **perform data-aware rule selection** for scoring preference data in the safety domain; Demonstrated **improved results on reward model benchmarks** using preference data scoring with the rule adaptor (**ICML 2025**)
- **Developed an automated method** to generate a large-scale, domain-specific dataset of graduate-level applied mathematics problems; **Benchmarked** leading closed- and open-source LLMs on this dataset and **performed in-depth error analysis**; Developed a framework to improve this domain specific ability through **tool usage** and **finetuning** (**NeurIPS MATH-AI 2024, ICLR 2025)**

Massachusetts Institute of Technology ***Cambridge, MA***
Undergraduate Researcher, Picower Institute *Sept.2017-May 2020*

- Conducted smFISH, IHC, q-PCR and behavioral assays to study the neural circuit for danger signal detection and avoidance during social behaviors and co-authored a paper published in ***Nature***

Undergraduate Researcher, McGovern Institute *Sept.2018-May 2020*

- Designed single-nanometer iron oxide nanoparticles as dopamine-responsive MRI sensors, developed brain-wide delivery methods to assess its distribution and functionality; Co-authored two papers published in ***JACS*** and ***PNAS***

PROFESSIONAL EXPERIENCE

Amazon, Research Intern *June. 2025-Sept. 2025*

- Created a novel benchmark for evaluating LLMs task performance considering both model capability and personalized preference alignment; in the application case of personalized recommendation, developed a process reward metric to balance both recommendation adoption and evidence faithfulness and leveraged it to perform RL post-training

Harvard AI Safety Student Team, Technical Fellow *Feb. 2025-May 2025*

Meta, Research Intern *May 2024-Aug. 2024*

- Developed a novel image-based feature representation tailored to high density sEMG and used customized CV models for gesture decoding and input feature attributions
- Introduced manifold capacity as a theoretical metric for representation quality evaluation and multimodal SSL loss
- Implemented generative models to extract disentangled factors in sEMG for generalization and data augmentation

Axoft, Software Intern *Sept. 2023-Dec. 2023*

- Developed and maintained in-house software pipelines for fluorescence imaging processing and spike sorting
- Applied statistical and machine learning models for neural decoding from population spiking and LFP data

PUBLICATIONS AND TALKS

The attention mechanism underlying relational object generation in text-to-image diffusion transformers. **New England Mechanistic Interpretability (NEMI) workshop**

Mapping a dynamic sensory panorama onto allocentric direction representations in goal-directed navigation. Selected talk at **Janelia Grounding Cognition in Mechanistic Insight Conference**

Mechanisms for balancing course stabilization and exploration. Talk at **Harvard Department of Neurobiology**.

Li, X., Chen X., Fan, J, Gao, M., Jiang, H. (2025). Entropy-aware Attribute Composition of Multi-head Reward Models (https://arxiv.org/abs/2503.20995). Submission to **AAAI 2026**

Li, X., Gao, M., Fan, J, Zhang, Z., Li, W. (2025). Data-adaptive Safety Rules for Training Reward Models (https://arxiv.org/pdf/2501.15453). **ICLR BiAlign 2025, ICML 2025**

Fan, J., Martinson, S., Wang, E.Y., Hausknecht, K. (2024). HARDMath: A Benchmark Dataset for Challenging Problems in Applied Mathematics (https://arxiv.org/pdf/2410.09988). **NeurIPS 2024 MATH-AI workshop, ICLR 2025**

Kwon, J.-T., Ryu, C., Lee, H., Sheffield, A., Fan, J., Cho, D. H., Bigler, S., Sullivan, H. A., Choe, H. K., Wickersham, I. R., Heiman, M., & Choi, G. B. (2021). An amygdala circuit that suppresses social engagement. **Nature**, 593(7857), 114–118.

Wei, H., Wiśniowska, A., Fan, J., Harvey, P., Li, Y., Wu, V., Hansen, E. C., Zhang, J., Kaul, M. G., Frey, A. M., Adam, G., Frenkel, A. I., Bawendi, M. G., & Jasanoff, A. (2021). Single-nanometer iron oxide nanoparticles as tissue-permeable MRI contrast agents. **Proceedings of the National Academy of Sciences**, 118(42).

Hsieh V., Okada S., Wei H., García-Álvarez I., Barandov A., Alvarado SR., Ohlendorf R., Fan J., Ortega A., Jasanoff A. (2019). Neurotransmitter-responsive nanosensors for T2-weighted magnetic resonance imaging. **Journal of the American Chemical Society**, 141 (40), 15751-15