

# Leyang Hu

✉ [leonleyanghu@gmail.com](mailto:leonleyanghu@gmail.com) |  [Google Scholar](#) |  [GitHub](#)

## RESEARCH INTERESTS

---

Deep Learning Theory, Self-supervised Learning (SSL), Representation Learning, Trustworthy AI, Large Language Models

## EDUCATION

---

### Brown University

Sep 2025 – Present

*Doctor of Philosophy in Computer Science (Advisor: Prof. Randall Balestriero)*

*Providence, RI*

Research Direction: Spline Theory of Deep Learning, Self-supervised Learning

### Brown University

Sep 2023 – May 2025

*Master of Science in Computer Science (AI / ML Research Track, GPA: 4.00 / 4.00)*

*Providence, RI*

Core Courses: Deep Learning, Research Topics in Self-supervised Learning, Robust Algorithms for ML,

Algorithmic Aspects of ML, Computational Methods for CogSci, Computational Linear Algebra

### University of Nottingham Ningbo China

Sep 2019 – May 2023

*Bachelor of Science in Computer Science with AI (First-Class Honors, GPA: 3.95 / 4.00)*

*Ningbo, China*

Core Courses: Machine Learning, AI Methods, Computer Vision, Language and Computation,

Mathematics for CS (Linear Algebra, Calculus, Probability, Statistics)

## RESEARCH EXPERIENCE

---

### Department of Computer Science, Brown University

Sep 2024 – Present

*Spline Theory of Deep Learning, Self-supervised Learning | Advisor: Prof. Randall Balestriero*

*Providence, RI*

- Showed that masked-image-modeling reconstruction loss underdetermines representation quality by constructing equal-loss encoders spanning 4%–74% linear-probe accuracy via a predictor–corrector walk along the loss level set; found conventional pretraining concentrates at 52%–73%, attributing downstream quality to optimization’s implicit bias.
- Led development of [stable-datasets](#), a unified data-loading package for deep learning that decouples dataset definitions from data hosting and provides switchable Arrow/Lance storage backends; used it to build an extensible benchmark evaluating six modern SSL methods across 41 image datasets, showing that modern SSL can be competitive with supervised learning for limited-data in-domain pretraining.
- Developed the first fine-tuning method ([Curvature Tuning](#)) that updates activation functions instead of model weights using insights from spline theory, and provably reshapes decision boundaries while empirically improving downstream accuracy by 10.20% and robustness by 2450.00% over LoRA with fewer than 60% of the parameters.
- Explained that large linear-probe performance variance in pretrained LLMs stems from misalignment between autoregressive pretraining features and downstream task features, and proposed [Next Token Perception Score](#) to quantify this alignment, showing strong correlation with linear-probe accuracy across 12 NLP datasets and eight LLMs (270M–8B parameters).

### School of Information Sciences, University of Illinois Urbana-Champaign

Mar 2024 – Oct 2024

*Large Language Model (LLM) | Advisor: Prof. Haohan Wang*

*Urbana, IL (Remote)*

- Led the development and experimental validation of [Revolve](#), an LLM self-refinement pipeline that improves optimization stability via discrete second-order differences, yielding gains of 7.8% in prompt optimization, 20.72% in solution refinement, and 29.17% in code optimization over TextGrad.
- Developed [DROJ](#), an embedding-level jailbreak attack that optimizes adversarial prompts and achieved a 98.46% success rate on AdvBench and MaliciousInstruct with Llama 2 7B.
- Conducted an in-depth [literature review](#) of jailbreak attacks and defenses for LLMs and vision-language models, synthesizing a unified framework to guide future research in large-model safety.

#### Carney Institute for Brain Science, Brown University

Nov 2023 – Mar 2024

Self-supervised Learning | Advisor: Prof. Thomas Serre

Providence, RI

- Developed a contrastive-learning model to align sEEG brain signals with speech audio, achieving 16.8% top-1 zero-shot retrieval accuracy for matching speech segments (vs. 0.56% chance).

#### Language Technologies Institute, Carnegie Mellon University

Nov 2021 – Jun 2024

Robustness | Advisor: Prof. Haohan Wang

Pittsburgh, PA (Remote)

- Contributed to [Robustar](#), an interactive toolbox for precise data annotation in robust vision learning, by integrating influence-function-based sample diagnosis, training and custom-upload APIs, and a client interface for streamlined end-user deployment.

## INDUSTRY EXPERIENCE

---

#### Yongjiang Laboratory

Jul 2022 – Apr 2023

Computer Vision | Advisor: Dr. Lijun Li

Ningbo, China

- Investigated depth-completion models for dense depth estimation in XR and implemented Sobel-edge-based loss weighting, improving edge-region F1 by 13% and accelerating training convergence by 5%.
- Identified out-of-distribution failure patterns and built an evaluation pipeline to amplify and quantify failure severity, which became a standard step in the lab's model assessment workflow.

## PUBLICATIONS AND PREPRINTS

---

- [1] **Hu, L.**, Gamba, M., Ghandikota, A., Nagaraj, A., & Balestrieri, R. (2026). Learning by reconstruction is an ill-defined prior for perception: A level set view. *Accepted at TAG-DS 2026 (Extended Abstract)*.
- [2] BuGhanem, S., **Hu, L.**, Zhang, H., Hajra, I., Nam, H., Facury, L., Vangaru, S., Sankar, P., Chakravarthy, A., Dhiman, A., Du, H., Merchant, H., Ghandikota, A., & Balestrieri, R. (2026). Is SSL ready for in-domain pretraining? A cross-dataset benchmark and analysis. *Submitted for peer review*.
- [3] **Hu, L.**, Gamba, M., & Balestrieri, R. (2025). Curvature tuning: Provable training-free model steering from a single parameter. *Accepted at NeurIPS 2025*. [arXiv:2502.07783](#).
- [4] Cheng, Y. A., **Hu, L.**, Huang, H., & Balestrieri, R. (2025). Next Token Perception Score: Analytical assessment of your LLM perception skills. *arXiv preprint*. [arXiv:2505.17169](#).
- [5] Zhang, P., Jin, H., **Hu, L.**, Li, X., Kang, L., Luo, M., Song, Y., & Wang, H. (2024). Revolve: Optimizing AI systems by tracking response evolution in textual optimization. *Accepted at ICML 2025*. [arXiv:2412.03092](#).
- [6] **Hu, L.**, & Wang, B. (2024). DROJ: A prompt-driven attack against large language models. *arXiv preprint*. [arXiv:2411.09125](#).
- [7] Jin, H., **Hu, L.**, Li, X., Zhang, P., Chen, C., Zhuang, J., & Wang, H. (2024). JailbreakZoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint*. [arXiv:2407.01599](#).

- [8] **Hu, L.**, Huang, Y., & Kar, P. (2022). An integrated single device framework for combined face recognition and COVID-19 detection using thermal infrared imagery. *Proceedings of the IEEE 22nd International Conference on Computer and Information Science (ICIS)*. DOI:10.1109/ICIS54925.2022.9882354.
- [9] Chen, C., Wang, H., **Hu, L.**, Zhang, Y., Lyu, S., Wu, J., Li, X., Sun, L., & Xing, E. P. (2022). Robustar: Interactive toolbox supporting precise data annotation for robust vision learning. *DataPerf Workshop at the 39th International Conference on Machine Learning (ICML 2022)*. *arXiv preprint*. [arXiv:2207.08944](https://arxiv.org/abs/2207.08944).

## TALKS AND PRESENTATIONS

---

**Curvature Tuning: Provable Training-Free Model Steering from a Single Parameter** **2025**  
*Oral Presentation at NECV 2025; Guest Lecture (Prof. Micah Goldblum, Columbia University)* *Multiple Venues*

**sEEG Decoding** **Jan 2024**  
*Tutorial at Brainstorm Challenge Workshop hosted by the Carney Institute for Brain Science* *Providence, RI*

## OUTREACH

---

**Department of Computer Science, University of Nottingham Ningbo China** **Jun 2023 – Aug 2023**  
*Coordinator | [Graduate Application Experience Handbook](#)* *Ningbo, China*

- Compiled the Graduate Application Experience Handbook by synthesizing graduate-application insights from 28 CS students; the resource received over 1,000 views and became widely used by junior students.

**Department of Computer Science, University of Nottingham Ningbo China** **Oct 2021 – Dec 2021**  
*Instructor | [Web Development Public Lecture Series](#)* *Ningbo, China*

- Designed and delivered a public web-development lecture series for 65+ students, covering frontend (HTML, CSS, JavaScript) and backend (Spring Boot) foundations to prepare students for full-stack engineering roles.

## SKILLS

---

**Programming Languages:** Python(PyTorch, TensorFlow, Transformers, Numpy), Java, C++, MATLAB, SQL, HTML/CSS/JS

**Languages:** English (Advanced; GRE: 332, TOEFL: 109), Chinese (Native)