Yiran Xu

Fudan University

yiranxu22@m.fudan.edu.cn github.com/Raizellll raizellll.github.io

Research Interests: Mechanistic Interpretability, Emergent Modularity, Representation Learning and Efficient ML.

EDUCATION

Fudan University **Sep. 2022 – Jun. 2026 (Expected)**
B.S. in Computer Science and Technology *Shanghai, China*

Relevant Coursework: Maching Learning, Natural Language Processing, Algorithms, Human–Computer Interaction

RESEARCH EXPERIENCE

Visiting Scholar — Post-hoc Modularity & Gradient-Flow Diagnostics **Jun. 2025 – Present**
EECS Department, University of Michigan *Supervisor: Prof. Robert P. Dick* *Ann Arbor, MI, USA*

- Established the ”Demand-Driven Modularity” theory: showed that input-distribution shifts do not induce modularity and that functional conflict is the key driver of physical parameter separation.
- Identified the ”Efficiency Bias” mechanism: Transformers maximize parameter reuse (high neuron overlap) and only separate into distinct manifolds under catastrophic functional interference.
- Reinterpreted the gradient starvation hypothesis by verifying early-layer optimality (L0 Probe Acc = 1.0), indicating that weight stagnation reflects feature sufficiency, not gradient loss.
- Preparing a first-author manuscript on the causal mechanism of modularity targeting ICML 2026.

Undergraduate Researcher — Neural Activation Analysis for LLM Evaluation **Sept. 2025 – Present**
Alex Reasoning Group, Fudan University *Supervisor: Prof. Yixin Cao* *Shanghai, China*

- Designed and implemented a latent-activation analysis pipeline to quantify reasoning depth, coherence, and creativity beyond standard accuracy metrics.
- Developed methods to extract interpretable low-rank activation subspaces and map semantic axes aligned with human-defined rubrics, connecting representation structure with multi-dimensional reasoning quality.

Undergraduate Researcher — Causal RL for Modular Reasoning in LLMs **Feb. 2025 – Jul. 2025**
MEMX Group, Fudan University *Supervisor: Prof. Li Shang* *Shanghai, China*

- Developed and validated a causal-RL framework for compact LLMs using MoE routing to disentangle decomposition, justification, and conclusion roles.
- Discovered and mitigated efficiency-bias collapse in self-training and introduced causal-consistency rewards that restored reasoning depth and stability across math, logic, and commonsense tasks.
- Contributed empirical findings that informed the later NAD interpretability framework.

INDUSTRY EXPERIENCE

Research Intern — LLM Reasoning & Code Generation **Jan. 2025 – Mar. 2025**
Huawei PaaS Lab *Mentor: Dr. Yuchi Ma* *Shenzhen, China*

- Designed a cognitive prompting pipeline for long-horizon code reasoning: decomposition → iterative synthesis → verification.
- Fine-tuned Qwen-2.5-72B on the TACO dataset (3.5k Codeforces problems) with 20-step reasoning trajectories, boosting symbolic planning accuracy.
- Analyzed reasoning traces to pinpoint bottlenecks and devised process-level correctness metrics.

HONORS AND AWARDS

Third Prize in China Mathematical Contest in Modeling (Top 15%, National)	**Nov. 2024**
Academic Excellence Scholarship of FDU	**Sept. 2024, Sept. 2023**

TECHNICAL SKILLS

Programming: Python (PyTorch, NumPy, Pandas), C++, SQL
ML/LLM Frameworks: HuggingFace Transformers, PEFT / LoRA, vLLM
Evaluation & Analysis: CKA / Representation Similarity, Clustering (K-means, UMAP), Activation Probing
Experiment & Infra: CUDA, Docker, Anaconda, Linux (tmux, JupyterLab)
Languages: Mandarin (Native), English (Fluent)