

YIRAN XU

Fudan University

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

Research Interests: Modular LLM architectures, mechanistic interpretability, and post-hoc modularization for reasoning.

EDUCATION

Fudan University

B.S. in Computer Science and Technology

Sep. 2022 – Jun. 2026 (Expected)

Shanghai, China

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

RESEARCH EXPERIENCE

Visiting Scholar — Post-hoc Modularization & Gradient-Flow Diagnostics

EECS Department, University of Michigan

Supervisor: Prof. Robert Dick

Jun. 2025 – Present

Ann Arbor, MI, USA

- Achieved 0.961 Macro-F1 with a post-hoc routed model, matching dense-model accuracy while increasing modularity.
- Diagnosed and demonstrated that late-layer collapse stemmed from update-magnitude imbalance rather than capacity limits.
- Introduced a gradient-flow analysis framework ($\Delta W/W$, CKA, routing metrics) that revealed how balanced gradients induce emergent modular structures.

Undergraduate Researcher — Neural Activation Analysis for LLM Evaluation

Alex Reasoning Group, Fudan University

Supervisor: Prof. Yixin Cao

Sept. 2025 – Present

Shanghai, China

- Designed and implemented a latent-activation analysis pipeline to quantify reasoning depth, coherence, and creativity beyond standard accuracy metrics.
- Devised methods to extract interpretable activation subspaces and mapped semantic axes aligned with human-defined rubrics.

Undergraduate Researcher — Causal RL for Modular Reasoning in LLMs

MEMX Group, Fudan University

Supervisor: Prof. Li Shang

Feb. 2025 – Jul. 2025

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

Huawei PaaS Lab

Mentor: Dr. Yuchi Ma

Jan. 2025 – Mar. 2025

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 Modelling (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)