# Jiayu Yang

Ph.D. Student, Information Hub, HKUST (GZ) | jyang729@connect.hkust-gz.edu.cn

## Education

**HKUST (Guangzhou)**, Information Hub — Sept 2024 – Present
M.Phil. and Ph.D. in AI | GPA 4.1/4.3 | Full Postgraduate Student Scholarship (PGS)

**Chongqing University**, HongShen Honor School — Sept 2020 – July 2024
B.Sc. in Statistics | Honors Degree (Top 5%)

- **Coursework:** Mathematical Analysis, Advanced Algebra, Differential Geometry, General Physics, Numerical Analysis

**University of California, Irvine**, Summer School — Summer 2022

## Research Interests

I study the internal mechanisms of large language models and turn that understanding into better post-training. Three threads run through my work.

- **Mechanistic interpretability of factual reasoning (TMLR, ACM MM 2025 Oral, ECCV 2026)**, spanning concept-level backdoor attacks, the first defense built against them, and continual learning over concept bottlenecks. In ACE **(ICLR 2026)** I showed that during multi-hop recall the implicit intermediate subject behaves as a "query neuron" that sequentially activates "value neurons" across transformer layers, and turned that account into a knowledge-editing method that controls these query–value pathways directly instead of overwriting facts blindly.
- **Reinforcement learning for latent reasoning.** In SWITCH **(EMNLP 2026)** I made on-policy RL well-defined over hidden-state recurrence by learning explicit boundary tokens, including synthesizing latent CoT for circumstance learning, which also left the latent computation open to probing and causal intervention rather than being an opaque block of continuous thought.
- **Multimodal understanding for coding tasks.** In InterpOCR **(ongoing)** I adapt MLLM mechanism tools to understand the codes in image forms, revealing how MLLM understanding the codes through OCR beyond various coding tasks.

What links them is a preference for explanations that are causal and testable over ones that are merely plausible. I want to bring that to **coding agents that are trained rather than prompted**: agentic RL on long-horizon software tasks, where rewards are sparse, trajectories are long, and credit assignment is the central difficulty, and where mechanistic analysis can say *why* a policy succeeds or fails, not only *that* it does.

## Research Experience

**Ph.D. Student**, HKUST (GZ), LAI Lab & LARK Lab — Sept 2026 – Present
Advised by Dr. Chengwei Qin & Dr. Zhijiang Guo

- LLM reasoning, mechanistic interpretability, and agents

**Research Intern**, Tencent Hunyuan — Feb 2026 – Present
Supervised by Han Hu

- LLM reasoning and world models

**M.Phil. Student**, HKUST (GZ), DI$^2$ Lab & LARK Lab — Sept 2024 – Present
Advised by Dr. Yutao Yue and Dr. Zhijiang Guo

- LLM reasoning and mechanistic interpretability

**Research Intern**, KAUST — Feb 2024 – June 2024
Supervised by Dr. Di Wang

- Explainable AI (XAI) and machine learning

## Publications

$*$ *denotes equal contribution.*

[1] **[ICLR 2026] ACE: Attribution-Controlled Knowledge Editing for Multi-hop Factual Recall**
**Jiayu Yang**, Yuxuan Fan, Songning Lai, Shengen Wu, Jiaqi Tang, Chun Kang, Zhijiang Guo, Yutao Yue | arXiv:2510.07896

[2] **[EMNLP 2026] Demystifying Hidden-State Recurrence: Switchable Latent Reasoning with On-Policy Reinforcement Learning**
**Jiayu Yang**, Chao Chen, Shengen Wu, Yinhong Liu, Yuxuan Fan, Lujundong Li, Songning Lai, Chengwei Qin, Zhijiang Guo | arXiv:2606.13106

[3] **[TMLR] CAT: Concept-Level Backdoor Attacks for Concept Bottleneck Models**
Songning Lai$*$, **Jiayu Yang**$*$, Yu Huang, Lijie Hu, Tianlang Xue, Zhangyi Hu, Jiaxu Li, Haicheng Liao, Yutao Yue | arXiv:2410.04823

[4] **[ECCV 2026] Dynamic-V2C: Editable and Continual Vision-to-Concept Bottleneck Models via Influence Functions**
Songning Lai∗, Shaofeng Liang∗, **Jiayu Yang**∗, Ninghui Feng, Yuxuan Fan, Wenshuo Chen

[5] **[TMLR, under review] Guarding the Gate: ConceptGuard Battles Concept-Level Backdoors in Concept Bottleneck Models**
**Jiayu Yang**, Yu Huang, Songning Lai, Gaoxiang Huang, Wenshuo Chen, Yutao Yue | arXiv:2411.16512

[6] **[Ongoing project] InterpOCR: How MLLM understanding CodeOCR?**
**Jiayu Yang**, Yuxuan Fan, Zhijiang Guo, Chengwei Qin |

[7] **[ACM MM 2025, Oral] Learning New Concepts, Remembering the Old: Continual Learning for Multimodal Concept Bottleneck Models**
Songning Lai, Mingqina Liao, Zhangyi Hu, **Jiayu Yang**, Wenshuo Chen, Hongru Xiao, Jianheng Tang, Haicheng Liao, Yutao Yue

[8] **[EMNLP 2026] Decomposing Visual Histories with Vision-Language Agents: Hierarchical Temporal Guidance for Compositional Image Generation**
Lujundong Li, Mingxu Zhang, Songning Lai, **Jiayu Yang**, Yuxuan Fan, Ziwei Xie, Feng Liu, Changwang Zhang, et al.

[9] **[Neurocomputing] A Comprehensive Review of Community Detection in Graphs** (IF 6.40, JCR Q1)
Jiakang Li, Songning Lai, Zhihao Shuai, Yuan Tan, Yifan Jia, Mianyang Yu, Zichen Song, Xiaokang Peng, Ziyang Xu, Yongxin Ni, Haifeng Qiu, **Jiayu Yang**, Yonggang Lu

[10] **[NeurIPS 2026, under review] Maintaining Informative Coherence: Migrating Hallucinations in Large Language Models via Absorbing Markov Chains**
Jiemin Wu, Songning Lai, Ruiqiang Xiao, Tianlang Xue, **Jiayu Yang**, Yutao Yue | arXiv:2410.20340

## Selected Projects

**SWITCH: Switchable Latent Reasoning with On-Policy Reinforcement Learning** **EMNLP 2026**

- Introduced a pair of learned boundary tokens (`<swi>`/`</swi>`) that make on-policy RL well-defined over Coconut-style hidden-state recurrence, while exposing the latent computation to direct mechanistic analysis.
- Built the full three-phase pipeline on Qwen3-8B — SFT for switch-position location, a latent curriculum, then Switch-GRPO (PPO-clipped, KL-anchored) with correctness, format, latent-usage, and brevity rewards — reaching **79.3%** on MATH-500 (**+25.7** over the strongest Coconut-style baseline at the same scale) and **89.2%** on GSM8K.
- Verified three mechanistic claims about the switch policy by linear probing (∼91.9% decodable from late hidden states), causal intervention, and logit lens; released the LoRA checkpoint and training set on HuggingFace.

**ACE: Attribution-Controlled Knowledge Editing for Multi-hop Factual Recall** **ICLR 2026**

- Traced the failure of prior knowledge editing on edits involving intermediate implicit subjects to an oversight of how chained knowledge is dynamically represented and used at the neuron level.
- Showed that during multi-hop reasoning implicit subjects act as "query neurons" that sequentially activate corresponding "value neurons" across transformer layers to accumulate information toward the final answer, and proved key metrics and properties through mathematical analysis.
- Outperformed state-of-the-art methods by **9.44%** on GPT-J and **37.46%** on Qwen3-8B, and showed that the semantic interpretability of value neurons is orchestrated by query-driven accumulation.

**CAT & ConceptGuard: Attacking and Defending Concept Bottleneck Models** **TMLR**

- Proposed **CAT** (**C**oncept-level Backdoor **AT**tacks), which injects triggers into conceptual representations during training to manipulate predictions precisely without compromising clean-data performance; **CAT+** adds a concept-correlation function that optimizes trigger–concept associations for higher attack effectiveness and stealth.
- Built **ConceptGuard**, the first defense tailored to concept-level backdoors: it clusters concepts by text distance and votes among classifiers trained on different concept subgroups, with theoretical guarantees below a trigger-size threshold while preserving CBM accuracy and interpretability.

## Awards

- **Honors Degree, Chongqing University** (Top 5%) — July 2024
- **Outstanding Student Scholarship, Chongqing University** — July 2022
- **MIT Outstanding Machine Learning Project** (Top 5%) — April 2022

## Academic Service

Reviewer for ICLR, ICML, NeurIPS, CVPR, ICCV, ECCV, and ICRA.