# Xiaotian YE

Beijing University of Posts and Telecommunications, Beijing, 100876, China
yexiaotian@bupt.edu.cn @ www.xiaotianye.me github.com/Acruxos

## Education

**B.Eng. in Computer Science** *Sep 2022 - Jun 2026*

*Beijing University of Posts and Telecommunications*

- **Overall GPA:** 92.32/100, 3.83/4.00
- **Activities and societies:** Member of ICPC Programming Team *(Gold medal in BUPT Campus Programming Contest for Freshmen, which is the team formation contest).*
- **Coursework:** Machine Learning (95), Design and Analysis of Algorithms (98), Data Structures (97), Python Programming (99), Matrix Theory (99), Foundation of Programming (98), Computer Systems (95), Formal Languages and Automata (96), Big Data Technology (97), Operating Systems (94), etc.

## Research Interests

Focus on the intersection of **knowledge mechanisms** and the **safety & trustworthiness** of foundation models (LLMs/VLMs), aiming to address three core research questions:

- How can we interpret the internal knowledge mechanisms of LLMs?
- How can we control and modify knowledge within LLMs to enhance trustworthiness and safety?
- How can we leverage such controllability to build safer and more trustworthy real-world AI systems?

## Publications & Manuscripts

- *Asterisk mark (*) denotes equal contribution as co-first author.*

### Research Highlights

[1] **LLM Unlearning Should Be Form-Independent**
Xiaotian Ye, Mengqi Zhang, Shu Wu
**IEEE S&P 2026**

[2] **Uncovering Overfitting in Large Language Model Editing**
Mengqi Zhang*, Xiaotian Ye*, Qiang Liu, Pengjie Ren, Shu Wu, Zhumin Chen
**ICLR 2025 Spotlight**

[3] **Knowledge Graph Enhanced Large Language Model Editing**
Mengqi Zhang*, Xiaotian Ye*, Qiang Liu, Pengjie Ren, Shu Wu, Zhumin Chen
**EMNLP 2024**

### Other Publications & Preprints

[4] **Monte-Carlo Interpretability Analysis on Knowledge Editing and Its Generalization Failure**
Xiaotian Ye, ChunYao Yang, Mengqi Zhang, Xiaohan Wang, Dongsheng Liu, Shu Wu
**Preprint, in submission**

[5] **Disentangling Knowledge Representations for Large Language Model Editing**
Mengqi Zhang*, Zisheng Zhou*, Xiaotian Ye, Zhaochun Ren, Zhumin Chen, Pengjie Ren
**ICLR 2026**

[6] **UIPE: Enhancing LLM Unlearning by Removing Knowledge Related to Forgetting Targets**
Wenyu wang*, Mengqi Zhang*, **Xiaotian Ye**, Zhaochun Ren, Zhumin Chen, Pengjie Ren
**EMNLP 2025 Findings**

[7] **KELE: Residual Knowledge Erasure for Enhanced Multi-hop Reasoning in Knowledge Editing**
Mengqi Zhang*, Bowen Fang*, Qiang Liu, **Xiaotian Ye**, Shu Wu, Pengjie Ren, Zhumin Chen et al.
**EMNLP 2025 Findings**

[8] **Towards Understanding the Effect of NTP Paradigm in Unstructured Knowledge Editing**
Zisheng Zhou, Mengqi Zhang, Shiguang Wu, **Xiaotian Ye**, Chi Zhang, Zhumin Chen et al.
**Preprint, in submission**

[9] **Spectral Characterization and Mitigation of Sequential Knowledge Editing Collapse**
Chi Zhang, Mengqi Zhang, **Xiaotian Ye**, Runxi Cheng, Zisheng Zhou, Pengjie Ren, Zhumin Chen
**Preprint, in submission**

[10] **Open Problems and a Hypothetical Path Forward in LLM Knowledge Paradigms**
**Xiaotian Ye**, Mengqi Zhang, Shu Wu
**Blogpost preprint, in submission**

## Internship / Research Experience

**Research Intern** *Jun 2023 – Present*

*Institute of Automation, Chinese Academy of Sciences* *Beijing, China*

◊ Work at NLPR & MAIS (State Key Lab of Multimodal AI Systems, formerly known as National Lab of Pattern Recognition; under Prof. Tieniu Tan's group), where I am supervised by Prof. Shu Wu and work with Dr. Mengqi Zhang; in collaboration with GAI Lab, Shandong University.

◊ **Project 1: Knowledge Editing and Unlearning of LLMs.** Contributed across the full research pipeline (survey, coding, experiments, rebuttal, and camera-ready), serving as a primary contributor for three papers: (1) proposed a knowledge graph–enhanced editing framework (EMNLP 2024), (2) first identified and analyzed overfitting phenomena in model editing (ICLR 2025 Spotlight), and (3) independently conducted research analyzing form-dependent bias in unlearning (S&P 2026, top security venue), receiving near-Distinguished Paper level scores (1,2,2,2; 1 is the highest; historical Distinguished typically 1,1,2,2) with a fully positive meta-review with no noteworthy concerns. Leading a multimodal knowledge editing project targeting ECCV 2026. Also contributed to & co-authored five additional papers (2*EMNLP 2025 Findings & 1*ICLR 2026 & 2 papers under review at ICML/ACL).

◊ **Project 2: Interpretability Analysis of LLM Knowledge Representation.** Leading a project that proposes a new class of interpretability tool by applying Monte Carlo-based sampling of latent representations and gradient-based analysis to study distributional properties of knowledge vectors and their solutions, explaining why representation-level control methods such as knowledge editing struggle to generalize. In submission to ICML 2026 as first author.

◊ **Project 3: Concept-Centric Safety Alignment for LLMs.** Leading a project that investigates concept-level alignment strategies encouraging models to align directly to harmful concepts rather than surface tokens, mitigating previously observed overfitting to literal lexical patterns and improving generalization under extreme OOD adversarial settings. Targeting submission to NeurIPS 2026 as first author.

## Selected Awards & Honors

- **CCF Elite Collegiate Award**, *annually award to 100 undergraduates nationwide, with each top-tier university nominating up to 4 candidates* *Sep 2025*
- International Silver Medal, *ICPC, Asia Regional Contest* *Dec 2023*
- National Silver Medal, *China Collegiate Programming Contest* *Oct 2023*

- National First Prize, *National English Competition for College Students, Final* *Jun 2023*
- National Individual Third Prize, *CCCC-GPLT, National Final* *May 2023*
- Merit Student, *Beijing University of Posts and Telecommunications* *Oct 2023*

## Skills

◊ **Programming** Especially experienced in Python, C++ and C; comfortable with JavaScript, Java. Experienced in programming under Unix-like environments including linux. Experienced in Competitive Programming.

◊ **Machine Learning** Experienced in PyTorch and NumPy; Have a good knowledge of theories and methods about machine learning and LLMs, especially in the field of LLM knowledge and interpretability, and understand common and important concepts in other domains as well.

◊ **English Proficiency** TOEFL BestScore 112 (R 29, L 30, S 25, W 28). Proficient in academic reading and writing.