# KEJIA CHEN

irinchan1222@gmail.com thecommonirin.github.io

## Education

**Zhejiang University** 09/2022 – Present
Ph.D. Candidate in Artificial Intelligence, VIPA Lab
*Advisors:* Prof. Zunlei Feng, Prof. Mingli Song

**Nanyang Technological University** 12/2025 – 07/2026
Visiting Student, *Advisor:* Prof. Tianwei Zhang

**East China Normal University** 09/2019 – 06/2022
M.Sc. in Software Engineering

**Nanjing University of Science and Technology** 09/2015 – 06/2019
B.Eng. in Software Engineering

## Research Interests

Foundation Model Safety Alignment, Reliable Agentic Systems.

## Publication Highlights

[1] **K. Chen**, J. Zheng, J. Zhang, M. Lin, J. Hu, Z. Feng, J. Lou, M. Song. "Token-Level Inference-Time Alignment for Vision-Language Models" in *Annual Meeting of the Association for Computational Linguistics* (**ACL**), 2026.
[2] **K. Chen**, J. Zhang, J. Yang, Z. Feng, M. Song. "Self-Improved Holistic Alignment for Preference Enhancement" in *Pattern Recognition* (**PR**), 2026.
[3] **K. Chen**, J. Zhang*, J. Hu, Y. Wang, Z. Feng, J. Lou, M. Song. "Assessing Safety Risks and Quantization-aware Safety Patching for Quantized Large Language Models" in *International Conference on Machine Learning* (**ICML**), 2025.
[4] **K. Chen**, J. Zhang, X. Liu, Z. Feng, X. Yang. "Private, Atomic, Incentive mechanism for Federated Learning based on Blockchain" in *International Conference on Blockchain and Trustworthy Systems* (**BCRA**), 2024.

## Co-Authored Publications

[1] J. Zhang, Y. Hu, **K. Chen**, L. He, J. Ma, J. Liu, J. Lou, X. Yang, R. Jia. "Understanding and Preserving Safety in Fine-Tuned LLMs" in *ACM Conference on Computer and Communications Security* (**CCS**), 2026.
[2] J. Zhang, L. He, **K. Chen**, J. Liu, J. Lou, X. Yang, R. Jia. "Safety at One Shot: Patching Fine-Tuned LLMs with A Single Instance" in *International Conference on Learning Representations* (**ICLR**), 2026.
[3] J. Zhang*, **K. Chen***, Z. Feng, J. Lou, J. Liu, M. Song, X. Yang. "Activation Approximations Can Incur Safety Vulnerabilities Even in Aligned LLMs: Comprehensive Analysis and Defense" in **USENIX Security Symposium**, 2025.
[4] J. Zhang, **K. Chen***, Z. Feng, J. Lou, J. Liu, M. Song, X. Yang. "SecPE: Secure Prompt Ensembling for Private and Robust LLMs" in *European Conference on Artificial Intelligence* (**ECAI**), 2024.
[5] J. Zhang, X. Yang, L. He, **K. Chen**, Y. Wang, J. Liu, X. Yang. "Secure transformer inference made non-interactive" in *Network and Distributed System Security Symposium* (**NDSS**), 2024.
*(Normalized Top-100 Security Papers.)*

## Selected Preprints

[1] **K. Chen**, J. Zhang, B. Li, P. Li, J. Lou, Z. Feng, M. Song, T. Zhang, R. Jia. "Mitigating Many-shot Jailbreak Attacks with One Single Demonstration", 2026.
[2] **K. Chen**, J. Zhang, Y. Du, J. Lou, Z. Feng, M. Song, R. Jia, T. Zhang. "Why Does Long-Context LLM Safety Break? A Positional Encoding Perspective", 2026.
[3] **K. Chen**, J. Zhang, T. Qiu, X. Pan, Z. Feng, M. Song, R. Jia. "Confidence-aware Alignment Makes Reasoning LLMs More Reliable", 2026.
[4] **K. Chen**, J. Zhang, L. He, Z. Feng, J. Liu, M. Song. "ARHarmGuide: Inference-Time Jailbreak Attacks through Token-Level Harmful Guidance", 2026.
[5] J. Zhang, **K. Chen**, J. Ma, Y. Hu, L. He, Y. Zhang, J. Liu, X. Yang, T. Zhang, R. Jia. "Beyond Similarity: Trustworthy Memory Search for Personal AI Agents", 2026.

## Internship Experience

**Alibaba Group Research Institute** 10/2025–Present
*Research Intern*
Aimed at improving fine-grained cross-modal synergy in Vision-Language Models. Designed a token-level feature alignment mechanism based on the Qwen3 architecture and streamlined the model training pipeline. Achieved simultaneous improvements in both core model benchmarks and overall computational efficiency.
*(Industrial Impact: The proposed explainable fine-grained perception agent has been successfully deployed in Taobao's core e-commerce content moderation system.)*

**Ant Group Research Institute** 02/2025–09/2025
*Research Intern*
Focused on advancing detection precision for complex visual manipulation and deepfake scenarios. Designed and optimized pixel-level forgery localization algorithms tailored for multimodal data. Won the Runner-up award (2nd place out of 500+ teams) in the IJCAI 2025 Multimodal Deepfake Fine-grained Localization Challenge.

**Qulian Technology Co., Ltd.** 07/2022–07/2024
*Research Intern*
Aimed at supporting social governance and risk monitoring applications through robust blockchain infrastructure. Participated in the deployment of HyperChain and independently built the foundational code framework. Developed and deployed the Common Infrastructure Platform for internal and external enterprise use.

## Academic Service

- **Conference Reviewer:** ICML, ICLR, NeurIPS, CVPR, ICCV, ECCV, ACL, EMNLP.
- **Journal Reviewer:** TIP, TIFS.

## Selected Awards & Honors

- 2nd Place Award (Global Runner-Up), IJCAI 2025 Multimodal Deepfake Fine-grained Localization Challenge
- Outstanding Graduate Innovation Achievement Award, Zhejiang Province
- Full Scholarship & Outstanding Student (Top 5%), Zhejiang University
- Full Scholarship & Outstanding Graduates, East China Normal University
- Outstanding Graduates, Nanjing University of Science and Technology