

Yuqi Chen

✉ yche1052@ucsc.edu

🐙 yuqichen4188.github.io

Research Interests

Model Interpretability, LLM Safety.

Education

- 2025 – Present 📖 **PhD, University of California Santa Cruz**, Baskin School of Engineering.
Advisor: Prof. Chenguang Wang
- 2024 – 2025 📖 **PhD, Washington University in St. Louis**, McKelvey School of Engineering.
Advisor: Prof. Chenguang Wang
- 2019 – 2022 📖 **Master, University of Chinese Academy of Sciences**, Aerospace Information Research Institute.
Advisor: Prof. Keming Chen
- 2015 – 2019 📖 **Bachelor, Xidian University**, School of Aerospace Science and Technology, Tsien Hsue-shen experimental class.
Major: Space Science and Technology

Publications

- [1] **Yuqi Chen**, Vincent Siu, Yang Liu, Dawn Song and Chenguang Wang. *Tunable Tool-Call Rates in LLM Agents via Representation Steering*, **Manuscript under review at AAAI 2027**.
- [2] X Liu, J Tu, **Y Chen**, S Xie, S Ren, T Shi, G Gantar, P. J. Gilbert, N Hynes, M Staver, W He, D Marn, A Low, X Zhang, D Miao, E Sandoval, D Lee, C Wang, W Guo, and D Song. *Agentified Agent Assessment Improves Standardization Across Heterogeneous Scenarios*, **Arxiv**.
- [3] **Y Chen**, V Siu, Y Liu, D Song, and C Wang. *Controlling Tool Use with Heading-Specific Activation Steering*. **Arxiv**.
- [4] K Montgomery, S Tan, **Y Chen**, S Zhuang, T Zhang, R Popa, and C Wang. *Budget-aware Test-time Scaling via Discriminative Verification*. **Manuscript under review at EACL 2027**.
- [5] Ye Yuan, Junyu Luo, Guancheng Wan, Jinsheng Huang, Chengwu Liu, Junwei Yang, Yifang Qin, Zhiping Xiao, Qingqing Long, Meng Xiao, Yiqiao Jin, Jianhong Tu, **Yuqi Chen**, Wei Ju, Zhongwei Wan, Yusheng Zhao, Xiao Luo, Yiwei Fu, Yizhou Sun, Wei Wang, Chenguang Wang, Ming Zhang. *A Comprehensive Survey of Evaluating Multimodal Foundation Models: Hierarchical Perspective and Extensive Applications*, **Preprint at SSRN**.
- [6] Y Potter, Z Wang, N Crispino, A Xiong, K Montgomery, F Pinto, E Chang, **Y Chen**, C Christodoulopoulos, M Ziyadi, R Gupta, C Wang, B Li, and D Song. *VMDT: Decoding the Trustworthiness of Video Foundation Models*, **Neural Information Processing Systems**, 2025.
- [7] Sijun Tan, Kyle Montgomery, and the rLLM Team. *rLLM v0.2: RL Training over General Agentic Programs*, **Project Blog**.
- [8] **Yuqi Chen**, Zequn Zhang, Guangyao Zhou, Xian Sun, Keming Chen. *Span-based dual-decoder framework for aspect sentiment triplet extraction*, **Neurocomputing**, Vol. 492, pp. 211–221, 2022.
- [9] **Yuqi Chen**, Keming Chen, Xian Sun, Zequn Zhang. *A Span-level Bidirectional Network for Aspect Sentiment Triplet Extraction*, In **Proceedings of EMNLP 2022** [Long Paper].

Open Source Projects

Key Contributor

📌 **rLLM, Reinforcement Learning for Language Agents.**

- The project has earned 5,100 GitHub stars.
- Blog: https://rllm-project.com/post.html?post=rllm_vo.2.md.
- GitHub: <https://github.com/rllm-org/rllm>.

Model Developer

📌 **Amazon Nova AI Challenge: Trusted Software Agents.**

- Developed LLM-based coding agents for secure multi-turn modification and extension of full-stack web applications.
- Focused on reliability, automated testing, and prevention of security vulnerabilities in AI-generated code.

Work Experience

2022 – 2023

📌 **Question Structuring for Automatic Problem-solving, Xizhi AI Education.**

Engaged in the construction of question modeling, including named entity recognition, relationship extraction, event extraction, and text classification. Improved the performance of question structuring by about 8%, and encapsulated the relevant algorithms for streamlined application.