

Zixuan Zhu

✉ zzx18742002@gmail.com  github.com/ternura143  ternura143.github.io

RESEARCH INTERESTS

- **Self-evolving agents:** how agents can continually and stably self-improve from their own experience.
- **Scalable evaluation:** how to scale agent evaluation efficiently and in parallel, and keep it robust and trustworthy.
- **Economics of AI:** how to reduce the token cost of agentic reasoning and get more capability from every token, until capable agents become viable on small, widely accessible models.

EDUCATION

Nanyang Technological University <i>M.Sc. in Financial Technology (Intelligent Process Automation)</i>	Aug. 2024 – Jun. 2026 Singapore, Singapore
Beijing University of Posts and Telecommunications <i>B.Mgmt. in E-Commerce Engineering with Law</i>	Sep. 2020 – Jun. 2024 Beijing, China
Queen Mary University of London <i>B.Eng. in E-Commerce Engineering with Law (dual degree)</i>	Sep. 2020 – May. 2024 London, UK

PUBLICATIONS

[Unified Context Evolution for LLM Agents](#)

Zixuan Zhu, Yitong Hu, Yong Dai, Junfeng Fang, Chunyang Jiang, Senkang Hu, Yuzhi Zhao
Under review [arXiv:2606.02304](#)

RESEARCH EXPERIENCE

Harbor-Index <i>Research Collaborator, Stanford University, advised by Prof. Ludwig Schmidt</i>	Dec. 2025 – Present Remote
• Built a compact 82-task agentic benchmark distilled from 6,627 candidate tasks across 54 benchmarks through difficulty filtering, automated auditing, and human review. • Led the token-analysis sub-team, studying agent failure modes and token efficiency across frontier and open models.	
CollabXplorer <i>Research Assistant, BUPT, advised by Prof. Yi Wang</i>	Sep. 2021 – May. 2023 Beijing, China
• Delivered scholar profiling, similarity search, and collaboration recommendation; recognized as a National Innovation & Entrepreneurship Project.	

INDUSTRIAL EXPERIENCE

Business Intelligence & Analysis Intern <i>SeaMoney & MariBank</i>	Jul. 2025 – Oct. 2025 Singapore, Singapore
• Built credit-risk models for retail-credit product lines across multiple markets, supported reject-rule governance and standardization, and automated liquidity and market-risk reporting.	
Data Science Intern <i>Beijing International Data Exchange</i>	Apr. 2023 – Dec. 2023 Beijing, China
• Developed fraud-detection and demand-forecasting models for insurance and operations scenarios.	

PROJECTS

Harbor ★2.7k A framework (by the Terminal-Bench team) for evaluating and optimizing agents and models in container environments. Built benchmark adapters (e.g., BFCL) and agent integrations (Qwen Code, Kimi CLI), developed the review bot, and led adapter review and standardization.	Core Maintainer
Transformers ★163k The model-definition framework for state-of-the-art machine learning models in text, vision, audio, and multimodal models, for both inference and training. Contributed the TIPSv2 (Text-Image Pre-training with Spatial awareness) family of contrastive vision-language encoders from Google DeepMind.	Contributor