

Hengyu Jin

✉ hengyuking@gmail.com | 🎓 | 🌐 | 🔗

EDUCATION

Tongji University (TJU)

Major: Bachelor of Engineering, Software Engineering | **GPA:** 4.60/5.00 (91.01/100)

Sep 2023 – Present

Research Interests: Reasoning and Reliability in Foundation Models; AI Agents

PUBLICATIONS & PREPRINTS

- **H. Jin**, S. Yang, and D. Wang. “Final Checkpoints Are Not Enough: Analyzing Latent Reasoning Faithfulness Along Training Trajectories.” *Under review at ACL Rolling Review*. [arXiv]
- W. Ye, X. Yuan, Y. Bin, **H. Jin**, L. Peng, P. Zeng, and H. T. Shen. “RISER: Orchestrating Latent Reasoning Skills for Adaptive Activation Steering.” *Findings of ACL 2026*. [Paper]
- T. Chen, X. Chen, Y. Shen, L. Zhang, **H. Jin**, and C. Cheng. “CORA: A Cognitive Reframing Dialogue Agent Powered by Large Language Models.” Accepted for publication in *IEEE Transactions on Affective Computing*. [DOI]
- J. Du, Y. Zhang, **H. Jin**, X. Zou, J. Zhu, and X. Fu. “Kairotask: Probing the Bridge Between Vague Intents and Spatiotemporal Contexts.” *CHI 2026*. [DOI]

RESEARCH EXPERIENCE

PRADA Lab, KAUST

Jul 2026 – Present

Visiting Student | Advisor: **Prof. Di Wang**

Saudi Arabia

| *Mechanistic Analysis of Latent Reasoning Faithfulness*

- Led a first-author investigation of how latent reasoning faithfulness evolves throughout training, showing why final-checkpoint performance alone can mask changes in a model’s internal reasoning process.
- Built a ProsQA evaluation pipeline combining BFS-verified counterfactual edits with latent-state interventions to test whether latent reasoning states causally reflect changes in problem structure across checkpoints.
- Compared latent and explicit reasoning approaches across diverse tasks, identifying distinct training dynamics that inform more reliable evaluation of reasoning systems.

| *Multi-User Agent Environment*

- Developing a verifiable interactive environment that generates scenario-grounded multi-user rollouts for training and evaluating multi principal agent systems, addressing the scarcity and high cost of collecting real-world multi-user interaction data.

Aether AI

Jun 2026 – Present

Research Intern | Advisor: **Prof. Biwei Huang** (UC San Diego)

Remote

| *Verifier for Robotics Agent System*

- Developing an execution-time verifier for a robotics agent system to detect failures during VLA execution and trigger early termination, replanning, or tool-assisted recovery.
- Built a natural failure-data generation pipeline in Genie Sim by identifying atomic subtasks from simulator states, checkpointing key execution states, and resuming rollouts with weaker VLA policies to produce realistic failure trajectories.
- Standardizing evaluation across existing verifier and reward-model benchmarks to support joint training on internally collected and public failure data.

Tongji University

Oct 2024 – Dec 2025

Research Intern | Advisor: **Prof. Ying Shen**

Shanghai

| *Hidden State Steering to Enhance LLM Latent Reasoning Skills*

- Co-developed a plug-and-play activation-steering framework that dynamically injects cognitive-primitive vectors into intermediate layers of frozen LLMs, achieving 3.4–6.5% average accuracy gains and 2–3× higher token efficiency than chain-of-thought baselines.
- Contributed to building a cognitive primitive construction pipeline based on Contrastive Activation Addition (CAA) + LLM-as-Judge filtering + KMeans clustering, distilling six categories of reusable reasoning vectors from Transformer hidden states.

| **CORA: CBT-Guided Post-Training for a Multi-Turn Dialogue Agent**

- Conducted LoRA-based supervised fine-tuning of Qwen2.5-7B-Instruct with LLaMA-Factory on CBT-guided cognitive-reframing dialogues for the CORA agent.
- Designed and implemented a Streamlit-based evaluation platform for blinded comparisons across model variants and configurable inference settings, supporting experimental validation and dialogue-data collection.

Tsinghua University

Jun 2025 – Oct 2025

Research Intern | Advisor: **Prof. Xinyi Fu**

Remote

| **Kairotask: Context-Aware Scheduling from Vague Intentions**

- To address the issue of disconnection between ambiguous intentions and execution scenarios, developed a backend reasoning architecture based on the large language model (Qwen-Plus) to achieve precise conversion from natural language to structured temporal metadata.
- Led the development of a complete mobile application by integrating iOS Core Location with geofencing technology, achieving location-based context-triggered notifications and reminders, and supporting an eight-day field deployment study.

SELECTED PROJECTS

Co-founder & Full Stack Development Lead | Follo AI: Context-Aware Scheduling Agent

Nov 2025 – Feb 2026

- Extended the Kairotask framework into a context-aware scheduling agent, designing a response–perception–memory architecture that combines LLM-based intent interpretation, environmental context, and structured user memory.
- Led end-to-end development of the mobile application and backend, delivering a functional MVP for proactive task management, context-aware recommendations, and real-world reminder triggering.

Team Lead & Core Developer | “Scrollscape” Vision Pro AR Application

Jun 2025 – Sep 2025

- Led the development of an immersive Chinese-garden education application for Apple Vision Pro, translating architectural and aesthetic concepts into interactive spatial-computing experiences.
- Implemented visionOS interaction and rendering components, including gesture-based navigation, virtual–physical scene transitions, and an automated aesthetic-evaluation module.

AWARDS AND HONORS

First Prize (East China Region), China Collegiate Computing Contest–Mobile App Innovation Division (2025)

First-grade Ethnic Class Special Scholarship, TJU (2023–2024, 2024–2025)

SKILLS

Relevant Coursework: Machine Learning, Computer Vision, Discrete Mathematics, Linear Algebra, Probability and Statistics

Languages: Chinese (Native); English (TOEFL: 4.5/6)

Programming: Python, C++, Swift

ML & Research Tools: PyTorch, Hugging Face Transformers, LLaMA-Factory

Development: SwiftUI, React, FastAPI

Systems & Tools: Git, L^AT_EX, Linux, Slurm