

Hefan Zhang

Hanover, NH • (603)3490048 • hefan.zhang.gr@dartmouth.edu

[LinkedIn](#) • [Homepage](#) • [Google Scholar](#)

I am a Ph.D. candidate in Computer Science at Dartmouth College, advised by Prof. Soroush Vosoughi. My research interests lie in natural language processing and machine learning, with a focus on LLM interpretability, reliability, and generation, as well as post-training LLM agents for long-horizon tasks. I am actively seeking internship opportunities focused on related fields.

EDUCATION

Dartmouth College, New Hampshire, USA

Sept 2023 - Now

Degree: PhD

Major: Computer Science

Shanghai Jiao Tong University (SJTU), Shanghai, China

Sept 2018 - Jun 2022

Degree: Bachelor of Engineering

Major: Information Engineering

Minor: Applied Mathematics

INTERNSHIP EXPERIENCE

Meituan | Large Language Model Research Intern

Jun 2026 - Sept 2026

- Addressed compounding planning, execution, and recovery errors in long-horizon tool-use agentic tasks, where a single policy entangles these behaviors and final-outcome rewards offer limited feedback on intermediate failures.
- Built HARP (Hierarchically Arranged Role Policies) with specialized Planner, Executor, and Repairer experts. Alternated RL with the Planner or Executor frozen, combining outcome rewards with verifier-based progress feedback.
- Routed experts and triggered repair using feedback; consolidated them via multi-teacher on-policy distillation (OPD).
- *In submission to ICLR 2027.*

RESEARCH EXPERIENCES

When Linguistic and Internal Confidence Diverge in Large Language Models | PhD Student Dec 2024 - Oct 2025

Advisor: **Prof. Soroush Vosoughi** (Associate professor of computer science at Dartmouth)

- Studied whether LLMs' self-reported confidence reflects their internal uncertainty, a key assumption in using verbal confidence for reliability assessment.
- Compared linguistic confidence with logits-based confidence and semantic entropy across 30 models, 8 classification tasks, and 2 generation tasks, evaluating association, magnitude agreement, and calibration under varied prompts.
- Found weak average instance-level association; instruction tuning often increased confidence gaps, while prompt changes shifted reported confidence without reliably improving calibration.
- *Accepted by EMNLP 2026 findings.*

Persuasion Simulation with Culture Backgrounds by LLM agents | PhD Student

Nov 2024 - Feb 2025

Advisor: **Prof. Soroush Vosoughi** (Associate professor of computer science at Dartmouth)

- Developed an LLM-driven framework to generate persuasion dialogues, simulating by modeling diverse personality and cultural profiles.
- Enhanced the diversity and realism of AI-generated persuasion dialogues by (1) reinforcing instructional prompting which ensures speaker characteristics guide every stage of dialogue generation and add a monitor component; (2) multilingual prompting, which aligns language use with speakers' native languages to better reflect cultural nuances.
- I was responsible for all the coding and experiments.
- *Accepted by EMNLP 2025 findings*

Persuasion Dataset Generation by LLMs | PhD Student

Jun 2024 - Oct 2024

Advisor: **Prof. Soroush Vosoughi** (Associate professor of computer science at Dartmouth)

- Developed a multi-agent LLM framework to overcome limitations in persuasive dialogue datasets, enabling scalable, high-quality generation for advancing research in computational and social sciences.
- Designed and implemented the entire code framework for a multi-agent large language model (LLM) system, enabling automated generation and annotation of persuasive dialogues for NLP research.
- Conducted extensive experimental evaluations, showcasing the framework's adaptability and effectiveness across diverse scenarios and persuasion strategies.
- I was responsible for all the coding and experiments.
- *Accepted by NAACL 2025 (Oral)*

PROJECT EXPERIENCES

Harbor Framework (Terminal-Bench Team) | Contributor Nov 2025 - Jan 2026

Harbor (terminal-bench) open-source evaluation framework for agents & LLMs

- Contributed to Harbor (terminal-bench) by developing the adapter for the PIXIU financial benchmark, enabling standardized evaluation of financial LLM tasks within Harbor's agent-based framework.
- Translated PIXIU tasks into Harbor's unified task format, including task specifications, instructions, environments, and evaluation logic, ensuring reproducibility and consistency with the original benchmark.
- Implemented automated adapter pipelines and conducted parity checks against the original PIXIU benchmark to validate correctness and execution fidelity.
- Documented and integrated the PIXIU adapter into the Harbor ecosystem, expanding Harbor's benchmark coverage to the financial domain.
- *Accepted by NeurIPS Dataset & Benchmark 2026*

SELECTED PUBLICATIONS

Hefan Zhang, Bingquan Zhang, Ming Cheng, Saeed Hassanpour, Weicheng Ma, and Soroush Vosoughi. "[When Linguistic and Internal Confidence Diverge in Large Language Models](#)." In Findings of the Association for Computational Linguistics: EMNLP 2026.

{**Hefan Zhang***, Weicheng Ma*}, Shiyu Ji, Farnoosh Hashemi, Qichao Wang, Ivory Yang, Joice Chen, Juanwen Pan, Michael Macy, Saeed Hassanpour, Soroush Vosoughi. "[Enhancing LLM-Based Persuasion Simulations with Cultural and Speaker-Specific Information](#)." In Findings of the Association for Computational Linguistics: EMNLP 2025.

Weicheng Ma, **Hefan Zhang**, Ivory Yang, Shiyu Ji, Joice Chen, Farnoosh Hashemi, Shubham Mohole, Ethan Gearey, Michael Macy, Saeed Hassanpour and Soroush Vosoughi. "[Communication makes perfect: Persuasion dataset construction via multi-LLM communication](#)." In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).

ACADEMIC SERVICE

Area Chair, ACL Rolling Review (ARR) - Oct 2025

Reviewer, ACL 2025, 2026

Reviewer, EMNLP 2025, 2026

Reviewer, NLPCC 2026

Reviewer, ICLR 2027

AWARDS, INTERESTS, SKILLS, AND COMMUNITY INVOLVEMENT

Selected Awards:

First Prize of The Chinese Mathematics Competitions	2019
Second Prize of Chinese Mathematical Olympiad in Senior	2017
Individual Silver Medal of World Mathematics Team Championship	2016
Second Prize of Chinese Mathematical Olympiad in Senior	2016

Skills:

Computer: Python, Matlab, C++, C#, Assembly, Latex, Microsoft offices (PowerPoint, Word, Excel)

Frameworks & Tools: Pytorch, Jupyter, Vscod, Linux, Git

Language: English (Fluent), Mandarin Chinese (Native)

Organization: President of Painting, Calligraphy and Seal Cutting Association (a club of Shanghai Jiao Tong University)