Zhang, Yang

Undergraduate Applicant for 2026 Fall PhD

stephanezhang85@gmail.com

EDUCATION

School of Electronics Engineering and Computer Science, Peking University, Beijing, China Aug. 2022 - Jul. 2026

Honors Program of B.E. in Intelligence Science and Technology (the **Zhi Class**) *Overall GPA:* 3.567/4.0

Core Courses: Algorithm Design and Analysis (Honor Track) (82) / Information Theory(93) / Practice of Programming in C&C++ (90) / Computer Vision (92) / Set and Graph Theory (87) ...

School of Computing, National University of Singapore, Singapore Aug. 2024 - Dec. 2024

Core Courses: Theory of Computation / Non-Linear Programming / Stochastic Process /… Exchange Student

PUBLICATIONS

[1] Qiu, T.*, Zhang, Y.*, Huang, X., Li, J., Ji, J., & Yang, Y. **ProgressGym: Alignment with a Millennium of Moral Progress.** ***Advances in Neural Information Processing Systems (NeurIPS) 2024.*** Spotlight. https://arxiv.org/abs/2406.20087

[2] Chen, Y.*, Zhao, Y.*, Zhang, Y.*, Zhang, A., Kawaguchi, K., Joty, S., Li, J., Chua, T.-S., Shieh, M. Q., & Zhang, W. **The Emergence of Abstract Thought in Large Language Models Beyond Any Language.** *Under Review for NeurIPS 2025.* https://arxiv.org/abs/2506.09890

RESEARCH EXPERIENCES

ProgressGym: Alignment with a Millennium of Moral Progress | Peking University | Co-First Author

Feb. 2024 - May 2024

Advisor: Yaodong Yang, Boya Assistant Professor at the Peking University Institute for Artificial Intelligence

- Introduced ProgressGym, a benchmark allowing alignment algorithms to learn mechanics of moral progress from history.
- Leveraged 9 centuries of historical text and 18 historical LLMs, and introduced 3 sub-tasks (tracking evolving values, anticipating moral progress, and regulating the feedback loop between human and AI values), enabling codification of real-world progress alignment challenges into concrete benchmarks.
- In response, presented extrapolative DPO\RLHF algorithms as baselines for future research, out-performing naive methods by up to 50%.

The Emergence of Abstract Thought in LLMs | National University of Singapore | Co-First Author

Dec. 2024 - May 2025

Advisor: Michael Qizhe Shieh, Assistant Professor at the Department of Computer Science of NUS

- Contributed the first framework to identify shared neurons supporting high-level reasoning across languages in large language models, providing evidence for abstract thought.
- Proposed a neuron-targeted training approach, improving reasoning tasks (GSM, MMLU) by up to 5% with less than 1% neurons trained using continual pre-training.

Cost-Aware Experimental Design Agent | UCSD | First Author

Jul. 2025 - Present

Advisor: Rose Yu, Assistant Professor at Department of Computer Science and Engineering of UCSD

- (Work in progress) Developing an agentic framework to optimize parameter tuning in costly experiments. Leverages large language models for knowledge and context-based reasoning while enhancing its previously lacking awareness on cost-efficiency.

Services and Activities

The 21nd "Ubiquant" programming competition, Peking University | Third Prize Apr. 2023

- Competed in teams of three in an ACM-styled programming competition, solved 7 out of 11 problems.

Interdisciplinary Contest in Modeling (ICM) | Honorable Mention (top 18%) Jan. 2020

- Solved Problem D: The Influence of Music using graph analysis.

Teaching Assistant, Introduction to Programming C at Peking University Feb. 2025 - June 2025

- Gave multiple lectures; prepared and supervised coding and lab homework.

Official Reviewer, ICML 2025 Mar. 2025

- Participated in the revision and rebuttal of three papers as an official reviewer in ICML 2025.

Skills

Solid experience in PyTorch implementations.

Large-scale experimenting: experiences in post-training and evaluating models with up to 70B parameters using multiple nodes.

Solid paper writing, presentation and rebuttal experience.

Utilizes: deepspeed, vllm, openRLHF, verl, sglang...

Standard English Tests: TOEFL: Total 115 (Reading 30, Listening 30, Speaking 28, Writing 27), GRE: V. 163, Q. 170, W. 3.5