Yang (Stephen) Zhang

https://ystephenzhang.github.io stephanezhang85@gmail.com ystephenzhang +86 15821531340

I work on alignment and domain tool-usage of LLMs to build AI that boosts social and intellectual advancement.

Education

Peking University Beijing, China
B.S. in Intelligence Science and Technology. *Overall GPA:* 3.562 / 4.0 2022.8–2026.6
the **Zhi Class**, Intelligence Science and Technology Honors Program (top 10% students)

National University of Singapore Singapore
Exchange Student at School of Computing. 2024.8 – 2024.12
Awarded with **Peking University's Singapore Exchange Scholarship**(only 4 recipients)

Publications

* Equal contribution † Corresponding author

[1] **Yang Zhang**, Yadi Cao, Sophia Sun, and Rose Yu†. "CAED-Agent: an Agentic Framework to Automate Simulation-Based Experimental Design". In: *Submitted to The Fourteenth International Conference on Learning Representations*. under review. 2025. URL: https://openreview.net/forum?id=nGihWDdQFI.

[2] Tianyi Qiu*, **Yang Zhang***, Xuchuan Huang, Jasmine Xinze Li, Jiaming Ji, and Yaodong Yang†. "ProgressGym: Alignment with a Millennium of Moral Progress". In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37. NeurIPS 2024 Spotlight. Curran Associates, Inc., 2024, pp. 14570–14607. DOI: 10.52202/079017-0465.

[3] Yuxin Chen*, Yiran Zhao*, **Yang Zhang**, An Zhang, Kenji Kawaguchi, Shafiq Joty, Junnan Li, Tat-Seng Chua, Michael Qizhe Shieh†, and Wenxuan Zhang. *The Emergence of Abstract Thought in Large Language Models Beyond Any Language*. NeurIPS 2025 Poster. 2025. arXiv: 2506.09890 [cs.CL]. URL: https://arxiv.org/abs/2506.09890.

Honors & Awards

- "Taotian" Scholarship (4 Recipients Department-wide), *School of Intelligence Science and Technology, Peking University* 2025
- Singapore Exchange Scholarship (4 Recipients), *Peking University* 2024

Experience

PKU Alignment Group, Peking University Beijing, China
Research Assistant under **Prof. Yaodong Yang** 2024.3 – 2024.8

- Worked on ProgressGym [1], a benchmark for alignment on progressing values. Leveraged 9 centuries of historical text, 18 historical LLMs, and 3 original sub-tasks to enable codification of real-world progress alignment challenges. Published at NeurIPS 2024 as Spotlight.
- Open-sourced PyPI package (200 monthly downloads), huggingface model collection (300 downloads in total) and huggingface leaderboard for ProgressGym.
- Participated in Safe-Sora [2], a safety alignment dataset on text-to-video generation.

TRAIL Lab, National University of Singapore Singapore
Exchange Student, Research Assistant under **Prof. Michael Shieh** 2024.10 – 2025.3

- Contributed a framework[3] to identify neurons supporting high-level reasoning across languages.
- Proposed a neuron-targeted training approach, improving reasoning tasks (GSM, MMLU) by up to 5% using continual-pretraining on less than 1% neurons, providing evidence of abstract thought. Published at NeurIPS 2025.

Rose Spatiotemporal Machine Learning Lab, UC San Diego La Jolla, California
Summer Research Intern under **Prof. Rose Yu** 2025.6 – 2025.9

- Developed an agent framework to solve cost-aware simulation configuration optimization, outperforming both Bayesian optimization and LLM baselines on 400 problems in three physics simulators. Manuscript under review for ICLR 2026.

Skills

English: TOEFL: Total 115 (Reading 30, Listening 30, Speaking 28, Writing 27)
Standard Tests: GRE: V. 163, Q. 170, W. 3.5
Implementation: Languages: Python, C, SQL. Libraries: PyTorch, deepspeed, vllm, verl, LangGraph
Large-scale experimentation: Post-training and evaluation of language models with up to 70B parameters; Deployment of tool-using LLM agent workflows.

Service

Departmental & University Service

- Executive Committee, *Students Union of Dept. of EECS, Peking University* 2024–2025
- Co-President, *Zhi Class, Peking University* 2022–2023
- Teaching Assistant, *Introduction to Computation (C)* 2025

Academic Service

- Official Reviewer, *ICML 2025, ICLR 2026* 2025
- Organizing Volunteer, *China 3DV 2025* 2025

[1] https://github.com/PKU-Alignment/ProgressGym
[2] Code open-source at: https://github.com/PKU-Alignment/safe-sora
[3] Code open-source at: https://github.com/ystephenzhang/multilingual_training.git