SHI QIU

(+1) 919-923-9983 Email: stephenqius@gmail.com ◇ Homepage ◇ Google Scholar

EDUCATION

Peking University 2022.09 - now
Undergraduate student in Astronomy, School of Physics

RESEARCH INTEREST

My research focuses on LLM reasoning and agentic AI. My research explores two fundamental directions: (1) Leveraging high-quality physics knowledge to enhance model reasoning beyond superficial pattern-matching, and (2) Developing and evaluating trustworthy agentic frameworks that demonstrate robust performance in real-world applications.

I'm holding a **U.S. permanent green card**, and are currently applying for phd positions.

SELECTED PUBLICATIONS

*: equal contributions. Full publication list is in Google Scholar.

[1] **Shi Qiu**, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, ..., Qing-Hong Cao, Hua Xing Zhu, PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models, in *NeurIPS 2025*. [2025]

[2] **Shi Qiu**, Zhun Wang, Tianneng Shi, Zhaorun Chen, Wenbo Guo, Dawn Song, AgentXploit: End-to-End Red-Teaming for AI Agents Powered by Multi-Agent Systems, *under review*. [2025]

[3] Peng Xia*, Siwei Han*, **Shi Qiu***, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, Huaxiu Yao, MMIE: Massive Multimodal Interleaved Comprehension Benchmark for Large Vision-Language Models, in *ICLR 2025 (Oral)*. [Paper] [Code] [2024]

[4] Yiyang Zhou*, Linjie Li*, **Shi Qiu***, Haibo Tong, Lijuan Wang, Huaxiu Yao, GLIMPSE: Do Large Vision-Language Models Truly Think With Videos or Just Glimpse at Them?, in *EMNLP 2025 Main Conference (Oral)*. [2025]

[5] Yifan Yao, Xeron Du, Bingli Wang, Kaijing Ma, Minghao Liu, **Shi Qiu**, ..., Jiaheng Liu, Stephen Huang, Ge Zhang SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines, in *NeurIPS 2025*. [Paper] [Website] [2025]

[6] Haibo Tong, Zhaoyang Wang, Zhaorun Chen, Haonian Ji, **Shi Qiu**, Siwei Han, Zhongkai Xue, Yiyang Zhou, Peng Xia, Kexin Geng, Mingyu Ding, Rafael Rafailov, Chelsea Finn, Huaxiu Yao MJ-VIDEO: Fine-Grained Benchmarking and Rewarding Video Preferences in Video Generation, in *NeurIPS 2025 (Spotlight)*. [Paper] [2025]

[7] Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, **Shi Qiu**, Shuai Wang, Jiamin Chen, Tiannan Wang, Jintian Zhang, Jing Chen, Wentao Zhang, Jiyu Chen, Ruipu Wu, Shiding Zhu, Xiangru Tang, Peng Cui, Huajun Chen, Ningyu Zhang, Mrinmaya Sachan, Agents: An Open-source Framework for Autonomous Language Agents, in *ICLR 2024 Workshop on Large Language Models for Agents.*[Paper], **Github 5.4k stars.**[project] [2023]

HONORS AND AWARDS

National Scholarship **2025**
Peking University Third Class Scholarship **2024**
Peking University Excellence in Research Award **2024**
Shu Qi Scholarship **2023**
Youth Award For Athletics **2023**

RESEARCH EXPERIENCES

Research Intern at Sunblaze Group at UC Berkeley 2025-Now
Mentor: Prof. Dawn Song

- Lead researcher of AgentXploit[2], an Open-Source, end-to-end agentic framework for automatic injection point exploit and validation.

Research Intern at AI4Physics Group at Peking University 2025-Now
Mentor: Prof. Hua Xing Zhu, Prof. Qing-Hong Cao and Prof. Ming-Xing Luo

- Lead researcher and initiator of PHYBench, coordinating 178 students from the School of Physics at Peking University to construct the first high-difficulty physics reasoning benchmark.
- Core contributor to IdeaSearch, dedicated to developing LLM-assisted solutions for optimization problems in physics, especially HEP, Astronomy and Fluid Dynamics.

Research Intern at UNC-Chapel Hill 2024-Now
Mentor: Prof. Huaxiu Yao

- Contributed to research on multimodal alignment and evaluation. Co-lead the project MMIE for constructing a novel benchmark for multimodal comprehension and reasoning, which contains 20,103 interleaved samples with mllm-as-as-judge scoring metric[3]. Proposed a benchmark called GLIMPSE to evaluate the perception capabilities of video LLMs[4]. Released a reward model for comprehensive evaluation of generated videos, in project MJ-Bench-Video[6].

Research Intern at AIWaves 2023
Mentor: Wangchunshu Zhou

- Participated in research of Self-evolving Agents and autonomous framework for multi-agent systems. Received 6k stars on Github repo.

Project Leader of CourseCommunity 2023-Now
Website: CourseCommunity

- Proposed an open-source, all-for-free platform providing university course notes and self-study guides. Received over 5,000 views and 2,000 downloads at our website.

EXTRACURRICULAR ACTIVITIES

Core Contributor and Community Ambassador of AlphaXiv 2024-Now

President of the Peking University Jump-Rope Team 2023-2024

- Beijing College Shuttlecock and Rope Skipping League: Men's 30-Second Double Under: First Place
- Mixed 1-on-1 30-Second Single Under: Second Place (Broke the event record)
- Consecutive Triple Under: Third Place

Class Monitor of Class 8 in the 22nd Undergraduate Cohort 2022-Now

ACADEMIC SKILLS

English Skills GRE **326** | TOEFL **112** | CET-6 **696**

Coding Skills **Python** | **LaTeX** | **Markdown**