Boxuan Li

boxuanli@microsoft.com | 412-251-1251 | https://li-boxuan.com

PUBLICATIONS

[1] Mike A Merrill, Alexander G Shaw, Nicholas Carlini, **Boxuan Li,** et al. (2026). Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. The Fourteenth International Conference on Learning Representations (ICLR 2026).
[2] Frank F. Xu*, Yufan Song*, **Boxuan Li***, et al. (2025). TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025). * stands for equal contribution.
[3] Aditya Bharat Soni, **Boxuan Li**, et al. (2025). Coding Agents with Multimodal Browsing are Generalist Problem Solvers. Findings of the Association for Computational Linguistics: EACL 2026.
[4] Xingyao Wang, **Boxuan Li**, et al. (2024). OpenHands: An Open Platform for AI Software Developers as Generalist Agents. The Thirteenth International Conference on Learning Representations (ICLR 2025).
[5] **Boxuan Li**, et al. (2020). MC-Explorer: Analyzing and Visualizing Motif-Cliques on Large Networks. 36th IEEE International Conference on Data Engineering (ICDE 2020) Demo Track.
[6] **Boxuan Li**, et al. (2020). Stable community structures and social exclusion. International Conference on Social Informatics.
[7] Cheng Wang, Xusheng Chen, Weiwei Jia, **Boxuan Li**, et al. (2018). 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18).

RESEARCH EXPERIENCE

Harbor | Cocreator — Sep 2025 – Present

Harbor is a framework for agent evaluation and training. It is a popular infrastructure software used by many AI research projects, including Terminal-Bench, SkillsBench, OTelBench, CompileBench, RuneBench, SWE-gen, OpenThoughts-Agent, endless-terminals, Ares, and SkyRL.

Terminal Bench | Maintainer — Jun 2025 – Present

Terminal Bench is a popular AI benchmark for terminal use, adopted by Claude, GPT, Gemini, etc. I was in the core team and in charge of both evaluation infrastructure and task review and auditing.

The Agent Company | Cocreator — Aug 2024 – May 2025

TheAgentCompany is the first benchmark that examines AI's ability to complete real-world consequential tasks. I designed a reproducible, extensible evaluation framework from scratch, led a team of 10+ software engineers to complete the coding part, and co-authored the paper.

OpenHands | Cocreator — Apr 2024 – Jan 2026

OpenHands is the most popular open-source coding agent. I am one of the early cofounders and an active maintainer of OpenHands. It has been downloaded by 1M+ times, reported by multiple media, and cited by 600+ academic papers.

JanusGraph | Maintainer — Sep 2019 – Jan 2026

Graph databases are a fundamental building block of AI applications to leverage the private domain knowledge. I served as Technical Steering Committee and led the development of JanusGraph, the most popular open-source distributed graph database. It is widely used by many companies including Microsoft, Apple, Netflix, Uber, RedHat, eBay, and Goldman Sachs. It has been downloaded by 1M+ times, and cited by 35+ academic papers.

WORK EXPERIENCE

Microsoft — Redmond, WA

Member of Technical Staff — Jan 2026 – Present

As a research engineer, I work on developing state-of-the-art large language models by improving its ability to "reason" about hard problems, especially math, stem, and coding questions.

Senior Software Engineer — Feb 2023 – Jan 2026

I work in Scope query optimizer team. Scope is the language for Microsoft's internal big data platform. Scope carries the company, processing Exabytes of data and nearly a million jobs every day. I lead the performance improvement efforts, including building sophisticated statistics model, tuning TPC-DS benchmarks, and developing state-of-the-art algorithms for join operations. My work has led to millions of dollars of savings per year for the company.

Goldman Sachs — Hong Kong

Software Engineer — Jul 2019 – Jul 2021

My team builds a massive live knowledge graph that enables company-wide business intelligence, machine learning, monitoring, and data governance. My contribution to JanusGraph project mostly happened during this period.

EDUCATION

Carnegie Mellon University, School of Computer Science
Master of Computational Data Science (Systems track), GPA: 4.0/4.0

- Teaching Assistant: Distributed Systems (15-640), Fall 2022

The University of Hong Kong
Bachelor of Engineering in Computer Science, GPA: 3.87/4.30

- Research Assistant: Systems & Database Group, 2018-2020