Yibo Peng

☎ 1-412-508-0360 ✉ yibopcmu@gmail.com in linkedin.com/in/yibo-peng-cs pppyb.github.io 🎓 Yibo Peng

Education

Carnegie Mellon University **Pittsburgh, PA**
Master of Science in Artificial Intelligence Engineering GPA: 3.83/4.0 *Aug 2023 – Dec 2024*

Beijing Jiaotong University & Lancaster University **Beijing, CN & Lancaster, UK**
*Bachelor of Science in Computer Science (**Honours**)* *Aug 2018 – July 2022*

Publications & Patents

- **Y. Peng***, J. Song*, L. Li*, R. Mangal, M. Christodorescu, C. Pasareanu, H. Zheng, B. Chen. “When ”Correct” Is Not Safe: Can We Trust Functionally Correct Patches Generated by Code Agents?” *International Conference on Learning Representations (ICLR)*, 2026 (Under Review). [arXiv]

- P. Xia*, **Y. Peng***, J. Wang*, K. Zeng, X. Wu, X. Tang, H. Zhu, Y. Li, S. Liu, Y. Lu, H. Yao. “MMedAgent-RL: Optimizing Multi-Agent Collaboration for Multimodal Medical Reasoning,” *International Conference on Learning Representations (ICLR)*, 2026 (Under Review). [arXiv]

- **Y. Peng**, Z. Wang, D. Fried. “Can Long-Context Language Models Solve Repository-Level Code Generation?” *LTI Student Research Symposium*, 2025 (Poster). [arXiv]

- B. Hu, “Intelligent Home Standards and Technologies.” *Human-Computer Interaction Application & Entertainment Type Equipment*. Ed. **Y. Peng**, Beijing: Tsinghua University Press, 2022. 144-155 & 212-247.

- Y. Tian, Z. Li, **Y. Peng**, 2021. Automatic control system and network of circulating water degassing devices. CN Patent Application 202120168813.7, filed January 2021.

Research Experience

Carnegie Mellon University **Pittsburgh, PA**
Adversarial Code Agent Research Advisor: Beidi Chen *May 2025 – Present*

- Designed and implemented **FCV-Attack**, a novel **black-box, single-query** attack that injects semantic, CWE-targeted suggestions into GitHub issue descriptions to induce code agents into generating patches that **pass all functional tests while embedding exploitable vulnerabilities**.
- Achieved an Attack Success Rate (ASR) of up to **56.3%** against industry-leading agent-model combinations, revealing a critical security blind spot in current code agent evaluation paradigms.
- Built a reproducible evaluation pipeline based on **SWE-bench** to systematically analyze the vulnerabilities of **12** leading code agent (e.g., SWE-Agent, OpenHands) and large language model (e.g., GPT, Claude) combinations.
- Demonstrated that the attack succeeds primarily by contaminating the model’s **internal state** (e.g., KV cache) rather than altering observable behaviors, proving the insufficiency of existing behavior-level defenses.

All Hands AI **Pittsburgh, PA**
Graduate Research Assistant Advisor: Graham Neubig *Feb 2025 – May 2025*

- Developed and implemented a semantic code search tool with RAG capabilities for the OpenHands agent framework, enabling AI agents to effectively search and utilize existing codebases.
- Built a complete RAG pipeline using sentence transformers and FAISS for efficient similarity search, supporting configurable embedding models and repository indexing with save/load functionality.

UNC & Microsoft Research **Chapel Hill, NC & Shanghai, CN (Remote)**
Research Intern Advisor: Huaxiu Yao *Jan 2025 – May 2025*

- Developed MMedAgent-RL, a reinforcement learning framework optimizing multi-agent collaboration for medical visual reasoning that simulates clinical GP → Specialist → GP workflows.
- Designed curriculum-based reinforcement learning strategy enabling attending physicians to progressively learn from specialist knowledge while addressing specialist inconsistencies.
- Achieved **state-of-the-art** performance across five medical VQA datasets, outperforming both proprietary models like GPT-4o and previous multi-agent systems by **20.7%** over SFT baselines.

Carnegie Mellon University **Pittsburgh, PA**
Repository-Level Code Generation Research Advisor: Daniel Fried *Aug 2024 – Jan 2025*

- Conducted a systematic comparison of Long-Context (LC) and Retrieval-Augmented Generation (RAG) approaches for repository-level code generation using CodeLlama-7B and Claude-3.5-sonnet.
- Discovered that **LC can outperform RAG for small, well-structured repositories (less than 40k tokens)**, while RAG remains superior for larger codebases with complex dependencies.
- Identified that **context organization** is more critical than chunking strategies, with semantic-based ordering significantly improving LC performance across all repository sizes.

Industry Experience

PricewaterhouseCoopers LLP (PwC) **Beijing, CN**
Development Engineer Intern – Quantitative Model Expert Team *Nov 2021 – Apr 2022*

- Developed a large VBA application to assess and calculate Expected Credit Loss (ECL) of accounts receivable.
- Reduced calculation time from **15 minutes to 10 seconds** by transitioning calculations to the database.
- Improved code efficiency by simplifying loops, reducing global variable usage, and optimizing function calls.
- Collaborated with cross-functional teams to integrate the model and over **230 listed companies** used it.