Yibo Peng

1-412-508-0360 yibopengcmu@gmail.com 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***, P. Xia*, 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," in Proceedings of the 14th International Conference on Learning Representations (**ICLR 2026**), Rio de Janeiro, Brazil, April 2026. [**Paper**]
- **Y. Peng**, J. Song, L. Li, X. Yang, M. Christodorescu, R. Mangal, C. Pasareanu, H. Zheng, B. Chen. "When *Correct* Is Not Safe: Can We Trust Functionally Correct Patches Generated by Code Agents?" in Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics *(**ACL 2026**)*. San Diego, California, July 2026. [**Paper**]
- **Y. Peng***, P. Xia*, D. Zhong*, K. Zeng, S. Han, Y. Zhou, J. Liu, R. Zhang, H. Yao. "SimpleOCR: Rendering Visualized Questions to Teach MLLMs to Read." in Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics *(**ACL 2026**)*. San Diego, California, July 2026. [**Paper**]
- **Y. Peng**, Z. Wang, D. Fried. "Can Long-Context Language Models Solve Repository-Level Code Generation?" *LTI Student Research Symposium*, 2025 (Poster). [**Paper**]
- **Y. Peng**, S. Chen, L. Lian, D. Wagner. "PIRO: On-Policy Distillation for Prompt Injection Robust Reasoning Models." *Submitted to **EMNLP** 2026.*
- S. Chen, **Y. Peng**, C. Sitawarin, D. Wagner. "SecPO: Principled Adversarial Training for Prompt Injection Security." *Submitted to **NeurIPS** 2026.*

Research Experience

University of California, Berkeley — **Berkeley, CA**
Adversarial Training for Prompt Injection Defense *Advisor: Prof. David Wagner* — *Mar 2025 – May 2025*

- Proposed **SecPO**, a DPO-based defense that approximates a non-existent secure oracle LLM by replacing the vulnerable reference distribution with the undefended model's clean-input distribution, eliminating the harmful KL pull toward insecure outputs.
- Built an efficient online adversarial training pipeline that runs an iterative attacker LLM inside the training loop, completing full training on LLaMA-3.1-8B in **18 hours on 16 H200 GPUs**.
- Reduced adaptive ASR by an order of magnitude on the strongest attacks (Genetic, TAP, PISmith Pass@800, GCG): from **95% (Meta-SecAlign) to 13%** on SEP, while preserving utility across 7 benchmarks.
- Scaled to Qwen3.6-27B via offline SecPO, achieving **7.5% AgentDojo Genetic ASR**—lower than GPT-5.5 (11.3%) and Gemini-3.1-Pro (90.0%)—despite no adversarial training on agentic tool calling, and raising the median attacker cost of residual successes by **18x** over the undefended model.

University of North Carolina at Chapel Hill — **Chapel Hill, NC(Remote)**
Visual Text Understanding in Vision-Language Models *Advisor: Prof. Huaxiu Yao* — *Sept 2025 – Nov 2025*

- Designed and validated a diagnostic framework that exposes OCR understanding gaps in VLMs by overlaying questions onto images, revealing **accuracy drops of 3-6%** across SOTA models (Qwen2.5-VL, GPT-4V).
- Developed **SimpleOCR**, a lightweight data augmentation technique that reduces diagnostic gaps from **3.4% to 0.8%**, demonstrating improved visual text comprehension through both supervised and reinforcement learning.

- Demonstrated plug-and-play applicability by integrating SimpleOCR into existing RL methods (GRPO, NoisyRollout), achieving **2-3% average improvements** with **2.3x larger gains on OCR-intensive benchmarks**.

Carnegie Mellon University **Pittsburgh, PA**
Adversarial Code Agent Research *Advisors: Prof. Beidi Chen & Prof. Corina Pasareanu* *May 2025 – Sept 2025*

- 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: Prof. 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.

Microsoft Research & UNC **Chapel Hill, NC & Shanghai, CN (Remote)**
Research Intern *Advisor: Dr. Yan Lu & Prof. 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: Prof. Daniel Fried* *Jan 2025 – April 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.

Services

Reviewer: NeurIPS 2025 Efficient Reasoning Workshop; NeurIPS 2025 ResponsibleFM Workshop; ICML 2025 R2-FM Workshop

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.