Yibo Peng

1-412-508-0360 | yibopengcmu@gmail.com | linkedin.com/in/yibo-peng-cs | pppyb.github.io | Google Scholar

EDUCATION

Carnegie Mellon University — **Pittsburgh, PA**
Master of Science in Artificial Intelligence Engineering — *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

[1] 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," in Proceedings of the 14th International Conference on Learning Representations (**ICLR 2026**), Rio de Janeiro, Brazil, April 2026. [**Paper**]

[2] **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**]

[3] **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**]

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

[5] **Y. Peng**, S. Chen, L. Lian, D. Wagner. "SecOPD: Mitigating Adaptive Prompt Injections by On-Policy Distillation." *Submitted to the Conference on Empirical Methods in Natural Language Processing* (**EMNLP 2026**).

[6] S. Chen, **Y. Peng**, C. Sitawarin, D. Wagner. "SecPO: Principled Adversarial Training for Prompt Injection Security." *Submitted to the Conference on Neural Information Processing Systems* (**NeurIPS 2026**).

RESEARCH EXPERIENCE

Berkeley Security Group, University of California, Berkeley *Researcher* *Feb 2026 – Present*

- Researched secure LLMs and model-level defenses against prompt injection attacks
- Researched adversarial post-training and preference optimization for LLM security
- Evaluated prompt-injection robustness for instruction-following and agentic tool-calling systems
- Advisor: David Wagner

InfiniAI Lab, Carnegie Mellon University *Researcher* *May 2025 – Present*

- Researched security and reliability risks in LLM-based code agents
- Researched vulnerability injection in functionally correct code patches
- Built evaluation pipelines for repository-level code repair and code-agent robustness
- Advisor: Beidi Chen

AIMINGLAB, UNC Chapel Hill *Researcher* *Oct 2025 – Jan 2026*

- Researched multimodal language models and visual text understanding
- Researched OCR-oriented training methods for improving visual grounding
- Evaluated multimodal reasoning models on OCR-intensive benchmarks
- Advisor: Huaxiu Yao

Microsoft Research *Researcher* *Feb 2025 – May 2025*

- Researched multi-agent systems for multimodal medical reasoning
- Researched reinforcement learning methods for coordinating medical reasoning agents
- Evaluated multimodal medical agents on visual question answering benchmarks
- Advisor: Yan Lu & Huaxiu Yao

Language Technologies Institute, Carnegie Mellon University *Researcher* *Jan 2025 – May 2025*

- Researched code LLMs for repository-level code generation and code understanding
- Researched retrieval-augmented generation and long-context modeling for coding tasks
- Built RAG-based semantic code search tools for LLM code agents
- Advisor: Daniel Fried & Graham Neubig

WORK 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.

SERVICES

Reviewer: ACM Computing Surveys, NeurIPS 2026, CVPR 2026, NeurIPS 2025 Efficient Reasoning Workshop; NeurIPS 2025 ResponsibleFM Workshop; ICML 2025 R2-FM Workshop