

# Pengyu Chang

pengyuch@andrew.cmu.edu | 412-670-4704 | linkedin.com/in/pengyuch | github.com/jmueducn

## Education

- 
- Carnegie Mellon University**, M.S. in Information Networking – Pittsburgh, PA Aug 2025 – May 2027
- **GPA:** 4.0/4.0
  - **Teaching Assistantship:** 11-485/685/785 Introduction to Deep Learning — mentored final projects, led recitations, and held office hours
- Shanghai Jiao Tong University**, B.E. in Software Engineering – Shanghai, China Aug 2021 – Jun 2025

## Experience

- 
- Software Engineering Intern**, Google – San Francisco, CA May 2026 – Present
- Building an end-to-end synthetic data generation pipeline for **Gemini** content safety, spanning multiple policy areas
  - Designed a two-stage approach that synthesizes structured policy **taxonomies**, then conditions generation on them for broad policy coverage and fine-grained control over the output distribution
  - Drove substantial gains in downstream safety classifier **F1** through iterative refinement of both stages, evaluated on accuracy, F1, and PR-AUC
- Software Engineer Intern**, Electronic Arts – Shanghai, China Jul 2024 – Aug 2024
- Integrated **Kiali** service mesh visualization into EADP's monitoring stack (**Grafana**, **Prometheus**) and cut its data volume by 85%, resolving memory bottlenecks and accelerating the DevOps team's error identification
- Research Assistant**, Shanghai Jiao Tong University – Shanghai, China Jun 2024 – Jun 2025
- Contributed to **LMTransplant** (EMNLP 2025), a novel data augmentation method for few-shot NLP: identified data-scarce downstream tasks, engineered prompts for synthetic data generation, and evaluated effectiveness across realistic augmentation scenarios
  - Designed and implemented **AdverTest** (first author, ISSTA 2026), a novel unit test generation framework for **Java** leveraging **multi-agent LLMs** via the **LangChain** framework to maximize fault detection capability
  - Improved fault detection rate on the Defects4J dataset by 8.56% over the best existing LLM-based methods and 63.30% over EvoSuite, while also increasing line and branch coverage

## Projects

- 
- CoLMbo-DF: Audio Language Model for Deepfake Detection paper** Oct 2025 – Mar 2026
- Led construction of **FakeReason**, a ~133K-pair audio dataset with **acoustic chain-of-thought** annotations for interpretable deepfake speech detection
  - Built a generation pipeline combining **ASVspoof 2019 LA** and **VoxCeleb2** with ~20K deepfakes synthesized via **Fish-Speech** and **CosyVoice2** TTS; used **Qwen3-30B** to produce CoT annotations grounded in prosodic, spectral, and physiological acoustic features
  - Dataset enabled a lightweight 1B-parameter **Llama 3.2** model to reach 98.7% accuracy on ASVspoof 2019, outperforming significantly larger audio language model baselines
- Minik8s: A mini container orchestration tool** Apr 2024 – Jun 2024
- Developed a management tool with functionality to add, query, modify, and delete API objects within the API Server using **GoLang**
  - Designed and implemented key control loops of Minik8s including **DeploymentController** and **HPAController** to manage container deployments and enable dynamic scaling
  - Implemented a persistent storage component based on **NFS**, including creation and management of Persistent Volumes (PV) and Persistent Volume Claims (PVC)
- LSM-KV: KVStore using Log-structured Merge Tree link** Apr 2023 – May 2023
- Built a key-value store in **C++** on a Log-Structured Merge-Tree architecture using **skip lists** and **Bloom filters**, spanning both in-memory and on-disk structures, and tuned throughput across LSM-Tree configurations

## Skills

---

**Languages:** Python, C++, C, Java, GoLang, JavaScript/TypeScript

**AI/ML & Infrastructure:** PyTorch, LangChain, LLM APIs, Kafka, Redis, MySQL, MongoDB, Prometheus, Grafana, Istio