# Zhiling Yan

Email: zhy423@lehigh.edu
Mobile: +1 (484) 705-6616

## Education

- **Lehigh University** — Pennsylvania, USA
  *Doctor of Philosophy in Computer Science; GPA: 3.95/4.00* — *Sep 2023 - Present*
  Advisor: Dr. Lichao Sun
- **Nanyang Technological University** — Singapore
  *Master of Science in Biomedical Data Science; GPA: 3.72/4.00* — *Aug 2022 - June 2023*
- **Sun Yat-sen University** — Guangdong, China
  *Bachelor of Science in Statistics; GPA: 3.64/4.00* — *Aug 2018 - June 2022*

## Research Interests

My research centers on agentic AI—developing self-evolving LLM agents and studying how agentic reasoning can earn humans' trust in high stacks domain—together with rigorous, contamination-free evaluation of medical and multilingual LLMs. Underlying these efforts is a broader interest in long-horizon agent optimization, where reinforcement learning and self-distillation turn sparse feedback into dense, reliable learning signals.

## Research Experience

- **Research Intern**: Harvard University — May 2026 - Aug 2026, USA
  Mentor: Dr. Quanzheng Li, Dr. Xiang Li
  - Developing recursion-structured on-policy self-distillation for optimizing long-horizon LLM agents.
- **Research Intern**: A*STAR - Agency for Science, Technology and Research — Oct 2022 - May 2023, Singapore
  Mentor: Dr. Weimiao Yu
  - Developed a quantitative solution to analyze EX02 expression pattern of different cell types in brightfield digital pathology IHC images, involving the development of a workflow to classify cell types.
- **Research Engineer Intern**: JD. Inc. — Jul 2021 - Dec 2021, China
  Mentor: Mr. Meng Chen
  - Developed an image2image model equipped with auxiliary classifiers to generate stylistic calligraphy from standard font. Ranked top 8 out of 500+ teams in JDHackthon competition. The paper was accepted by IJCAI 2022.

## Publications

**Google Scholar**: **4,344** citations | h-index: **12** | i10-index: **15** (* denotes equal contribution)

1. **Zhiling Yan***, Dingjie Song*, Hanrong Zhang, Wei Liang, Yuxuan Zhang, Yutong Dai, Lifang He, Philip S. Yu, Ran Xu, Xiang Li, Lichao Sun. "*OpenSkill: Open-World Self-Evolution for LLM Agents*", The 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2026. [Paper] [Code]
2. **Zhiling Yan**, Dingjie Song, Zhe Fang, Yisheng Ji, Xiang Li, Quanzheng Li, Lichao Sun. "*LiveMedBench: A Contamination-Free Medical Benchmark for LLMs with Automated Rubric Evaluation*", ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2026. [Paper] [Code]
3. **Zhiling Yan**, Zhe Fang, David J. King, Ann Pongsakul, Eashan Adhikarla, Hui Ren, Sunyang Fu, Quanzheng Li, Lifang He, Xiang Li, Hongfang Liu, Yonghui Wu, Lichao Sun. "*Agentic AI Enhances Physician Trust in Clinical Decision Making*", American Medical Informatics Association (AMIA) Annual Symposium, 2026. [Paper] [Code]
4. **Zhiling Yan**, Rong Zhou, Kai Zhang, Wenting Chen, Yuteng Zeng, Yi Luo, Zhe Fang, Hui Ren, Quanzheng Li, Yixuan Yuan, Tianming Liu, Carl Yang, Yong Chen, James Zou, Lifang He, Xiang Li, Lichao Sun. "*Evaluating GPT-4V for Multi-modal Medical Applications: A Comparative Study Using Open-source Datasets*", IEEE Transactions on Radiation and Plasma Medical Sciences (TRPMS), 2026. [Paper] [Code]
5. **Zhiling Yan**, Sifan Song, Dingjie Song, Yiwei Li, Rong Zhou, Weixiang Sun, Zhennong Chen, Sekeun Kim, Hui Ren, Tianming Liu, Quanzheng Li, Xiang Li, Lifang He, Lichao Sun. "*SAMed-2: Selective Memory Enhanced Medical Segment Anything Model*", International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2025. [Paper] [Code]
6. Tianyi Zhang*, **Zhiling Yan***, Chunhui Li, Nan Ying, Yanli Lei, Yunlu Feng, Yu Zhao, Guanglei Zhang. "*CellMix: A General Instance Relationship based Method for Data Augmentation Towards Pathology Image Classification*", IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 2025. [Paper] [Code]
7. Rong Zhou*, Zhengqing Yuan*, **Zhiling Yan***, Weixiang Sun*, Kai Zhang, Yiwei Li, Xiang Li, Quanzheng Li, Lifang He, Lichao Sun. "*TTT-UNet: Enhancing U-Net with Test-Time Training Layers for Biomedical Image Segmentation*", Medical Imaging with Deep Learning (MIDL), 2026. (Oral) [Paper] [Code]
8. Kai Zhang, Rong Zhou, Eashan Adhikarla, **Zhiling Yan**, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, Lichao Sun. "*A generalist vision–language foundation model for diverse biomedical tasks*", Nature Medicine, 2024. [Paper] [Code]
9. Yihan Cao, Siyu Li, Yixin Liu, **Zhiling Yan**, Yutong Dai, Philip Yu, Lichao Sun. "*A Survey of AI-Generated Content (AIGC)*", ACM Computing Surveys, 2025. [Paper]

10. Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, **Zhiling Yan**, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, Tianming Liu, Pheng-Ann Heng, Quanzheng Li. "*MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation*", Medical Image Analysis, 2024. [Paper] [Code]

## Projects & Engineering

- **OpenSkill: Open-World Self-Evolution for LLM Agents**: we study open-world self-evolution. An agent builds skills and its own verification signals from open-world resources with no target-task supervision. [EMNLP'26] **[Co-first Author]**
  - **Motivations:** Existing self-evolving agents assume a usable learning loop—curated skills, successful traces, or verifier signals—yet real open-world deployments may provide only a task prompt, with none of these.
  - **Methodology:** We propose OpenSkill, a three-stage pipeline that acquires grounded knowledge and verification anchors from documentation, repositories, and the web, evolves skills against self-built virtual tests grounded in those anchors rather than target answers, and deploys them zero-shot—keeping target-task supervision out of skill construction.
  - **Results:** OpenSkill attains the best automated pass rate across three benchmarks and two model families (e.g., +8.9/+8.8 over the strongest closed-world baseline on SkillsBench), transfers across models without adaptation, and self-builds a verifier covering 88.9% of ground-truth test intents.
- **LiveMedBench: A Contamination-Free Medical Benchmark for LLMs with Automated Rubric Evaluation**: we build a system that automatically harvests real-world clinical data and constructs objective rubrics for scalable, contamination-free medical LLM evaluation. [KDD'26] **[First Author]**
  - **Motivations:** Existing medical benchmarks are static—prone to data contamination and temporal misalignment—and rely on shallow lexical-overlap metrics or subjective LLM-as-a-Judge scoring, neither of which objectively verifies clinical correctness.
  - **Methodology:** We build an end-to-end system that weekly mines fresh clinical cases from verified online communities, a Multi-Agent Clinical Curation Framework that denoises and validates cases against evidence-based medical standards, and an Automated Rubric-based Evaluation Framework that decomposes physician responses into granular case-specific criteria for objective grading; the benchmark is a byproduct of this pipeline.
  - **Results:** The pipeline yields a continuously refreshed benchmark of 2,756 cases across 38 specialties with 16,702 rubric criteria; the automated rubric aligns better with physician experts than LLM-as-a-Judge, and evaluation reveals the best model reaches only 39.2% with 84% of models degrading on post-cutoff cases.
- **SAMed-2: Selective memory enhanced medical segment anything model**: we introduced a new foundation model for medical image segmentation. [MICCAI'25] **[First Author]**
  - **Motivations:** Adapting SAM directly to medical domain remains challenging due to the complexity of medical data, nois annotations, and continual learning requirements across diverse modalities and anatomical structures.
  - **Methodology:** we introduce a temporal adapter into the image encoder to capture image correlations and a confidence-driven memory mechanism to store high-certainty features for later retrieval.
  - **Results:** SAMed-2 achieves state-of-the-art performance on both internal and external tasks, notably improving external zero-shot results by 10.53%.
- **CellMix: Instance Relationship based Method for Data Augmentation Towards Pathology Image Classification**: We identify the domain characteristics and purtabation of pathology images and introduce curriculum learning with extra shuffling strategy for better pathology image modeling. [TNNLS'25] **[Co-first Author]**
  - **Motivations:** Existing data augmentation methods fail to capture the domain characteristics of pathology images including local specificity, global distribution features and inner/outer-sample instance relationship. And it is much tougher to do the augmentation on pathology images considering more perturbation and higher inconsistency in the distribution of the same label.
  - **Methodology:** First, we propose a CellMix framework based on an in-place shuffle strategy, which regroups the patches in pathology images and generates augmented samples. Secondly, we design a fix-position ratio scheduler and a patch-size scheduler towards the in-place shuffle process. We continuously change the patch size and the fixed-position ratio of pathology images to reduce the difficulty of the training task.
  - **Results:** On various pathology image classification benchmarks, CellMix significantly outperforms all other mixing-based augmentation methods. CellMix boosts the performance of ViT-base model up to 4.59%.

## Honors and Awards

- *Best Research Award*, Department of Computer Science and Engineering, Lehigh University — May, 2026
- NSF ACCESS Award (Award CIS260542) — Apr, 2026
- Lehigh University Fellowship — Sep, 2025
- Lehigh University Fellowship — Sep, 2023
- *Finalist* at the 7th JD Hackthon Competition — Nov, 2021
- *Provincial Second Prize* at Contemporary Undergraduate Mathematical Contest in Modeling — Sep, 2019

## Skills/Qualification/Others

- **Languages**: Python, R, MATLAB, C++
- **Frameworks**: Pytorch, Tensorflow, Keras
- **Tools**: MySQL, GIT
- **LLM Training & Post-training**: verl, LLaMA-Factory, FSDP, DeepSpeed, FlashAttention
- **Inference & Serving**: vLLM, litellm
- **Infrastructure**: Docker, Slurm
- **Academic Service**: Reviewer for NeurIPS, KDD, COLM, MICCAI, ICHI, IEEE TMI, and IEEE TAI