Yifan Yang

Phone: 301-323-5740
Email: yifan.yang3@nih.gov · yang7832@umd.edu

Areas of Interest

Medical AI · AI safety · Natural Language Processing · Machine Learning.

Education

Current **PhD Candidate in Computer Science**,
University of Maryland College Park.
Advisor: Furong Huang

Current **NIH Visiting fellow**,
Advisor: Zhiyong Lu, PhD FACMI FIAHSI

2020 BS in Computer Science with high honors,
University of Maryland College Park.

Skills

Programming: Python, R, C, C++, Java
Frameworks: transformers, accelerate, deepspeed, pytorch, numpy, pandas
Data Management: MySQL

Awards

2022 - Current NIH Predoctoral Visiting Program Award

2024 NIH director's challenge award

Projects

Medical LLM Safety
As LLMs become increasingly integrated into clinical and medical workflows, we investigate their safety, fairness, and robustness in real-world settings. This line of works focuses on identifying potential biases in medical AI, such as those stemming from race, gender, imaging modalities, or socioeconomic status, and how these can impact model behavior. We explore how adversarial inputs can manipulate LLM outputs in clinical scenarios and propose defense strategies to enhance model reliability. To address privacy concerns, we develop a method that functions as an inference-time guard, preventing sensitive health information leakage from LLMs trained on domain-specific corpora. To comprehensively evaluate safety in medical applications, we introduce MedGuard, an expert-verified benchmark dataset designed around five core safety principles and ten clinically relevant aspects, grounded in practical use cases rather than adversarial prompts.

Contextual-Augmented LLMs
While state-of-the-art LLMs exhibit impressive reading comprehension and reasoning, their performance can degrade in specialized domains where training data is sparse. To bridge this gap in

medical and biomedical applications, our work combines RAG with targeted foundation-model fine-tuning. We built a pipeline that lets LLMs generate and invoke medical calculators, automatically extracting patient data from clinical notes to compute risk scores. To bolster multi-hop reasoning, we trained the model to formulate more precise, cascading queries, improving its ability to connect disparate pieces of information. Finally, we enhanced gene-centric reasoning by fine-tuning on curated genomic databases and applying reinforcement learning from AI feedback, yielding more accurate and explainable gene-related inferences.

Assisting Clinical Tasks through LLMs

To accelerate and streamline patient recruitment for clinical research, we developed TrialGPT, an end-to-end LLM-powered system for zero-shot patient-to-trial matching that is in the experimental stage at the NIH. TrialGPT's three-stage pipeline first retrieves a focused subset of relevant studies from a large trial corpus, then evaluates eligibility at the criterion level using fine-tuned LLMs, providing transparent, faithful explanations, and finally ranks the matched trials to surface the best candidates. By integrating retrieval-augmented generation with task-specific fine-tuning and validation against real patient data, TrialGPT significantly reduces manual screening effort and improves both recall and precision in trial matching.

Publications

2025 **Beyond Multiple-Choice Accuracy: Real-World Challenges of Implementing Large Language Models in Healthcare.** Yifan Yang, Qiao Jin, Qingqing Zhu, Zhizheng Wang, Francisco Erramuspe Álvarez, Nicholas Wan, Benjamin Hou, and Zhiyong Lu. *Annual review of biomedical data science, 10.1146/annurev-biodatasci-103123-094851*

2025 **Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey.** Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, Furong Huang. *NAACL findings, 2025*

2024 **Matching patients to clinical trials with large language models.** Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, Zhiyong Lu. *Nature Communications volume 15, Article number: 9074 (2024)*

2024 **Unmasking and Quantifying Racial Bias of Large Language Models in Medical Report Generation.** Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, Zhiyong Lu. *Communications Medicine, volume 4, Article number: 176 (2024)*

2024 **Opportunities and challenges for ChatGPT and large language models in biomedicine and health.** Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, Zhiyong Lu. *Briefings in Bioinformatics, Volume 25, Issue 1, January 2024, bbad493*

2024 **A survey of recent methods for addressing AI fairness and bias in biomedicine.** Yifan Yang, Mingquan Lin, Han Zhao, Yifan Peng, Furong Huang, Zhiyong Lu. *Journal of Biomedical Informatics 154, 104646.*

2024 **GeneGPT: augmenting large language models with domain tools for improved access to biomedical information.** Qiao Jin, Yifan Yang, Qingyu Chen, Zhiyong Lu. *Bioinformatics, Volume 40, Issue 2, February 2024, btae075.*

2023 **Improving model fairness in image-based computer-aided diagnosis.** Mingquan Lin, Tianhao Li, Yifan Yang, Gregory Holste, Ying Ding, Sarah H. Van Tassel, Kyle Kovacs, Zhangyang Wang, Zhiyong Lu, Fei Wang, Yifan Peng. *Nature Communications 14.1 (2023): 6261.*

2023 **C-Disentanglement: Discovering Causally-Independent Generative Factors under an Inductive Bias of Confounder.** Xiaoyu Liu, Jiaxin Yuan, Bang An, Yuancheng Xu, Yifan Yang,

Furong Huang. *NeurIPS 2023*

2022 **Comfetch: Federated Learning of Large Networks on Memory-Constrained Clients via Sketching.** Tahseen Rabbani, Brandon Feng, Yifan Yang, Arjun Rajkumar, Furong Huang. *AAAI-22*

2020 **Epiviz File Server: Query, Transform and Interactively Explore Data from Indexed Genomic Files.** Jayaram Kancherla, Yifan Yang, Hyeyun Chae, Hector Corrada Bravo. *Bioinformatics*
Poster & Presentation at ISMB 2019.

Under review at journal or conference (Avaliable as preprints)

2025 **CT-Bench: A Comprehensive Benchmark for Multimodal AI in Computed Tomography Analysis.** Qingqing Zhu, Qiao Jin, Tejas Sudharshan Mathai, Yin Fang, ZhiZheng Wang, Yifan Yang, Maame Sarfo-Gyamfi, Benjamin Hou, Ran Gu, Praveen T. S. Balamuralikrishna, Ken-
2025 neth C. Wang, Ronald Summers, Zhiyong Lu **RAG-Gym: Optimizing Reasoning and Search Agents with Process Supervision.** Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, Zhiyong Lu, Aidong Zhang.

2024 **Memorization in Large Language Models in Medicine: Prevalence, Characteristics, and Clinical Implications.** Anran Li, Mengmeng Du, Yu Yin, Yan Hu, Zihao Sun, Yihang Fu, Lingfei Qian, Erica Stutz, Xuguang Ai, Qianqian Xie, Rui Zhu, Jimin Huang, Yifan Yang, Siru Liu, Yih-Chung Tham, Lucila Ohno-Machado, Hyunghoon Cho, Zhiyong Lu, Hua Xu, Qingyu Chen.

2024 **Adversarial Attacks on Large Language Models in Medicine.** Yifan Yang, Qiao Jin, Furong Huang, Zhiyong Lu.

2024 **Ensuring Safety and Trust: Analyzing the Risks of Large Language Models in Medicine.** Yifan Yang, Qiao Jin, Robert Leaman, Xiaoyu Liu, Guangzhi Xiong, Maame Sarfo-Gyamfi, Changlin Gong, Santiago Ferrière-Steinert, W. John Wilbur, Xiaojun Li, Jiaxin Yuan, Bang An, Kelvin S. Castro, Francisco Erramuspe Álvarez, Matías Stockle, Aidong Zhang, Furong Huang, and Zhiyong Lu.

2024 **AgentMD: Empowering Language Agents for Risk Prediction with Large-Scale Clinical Tool Learning.** Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, Zhiyong Lu.

2024 **Knowledge-guided contextual gene set analysis with large language models.** Zhizheng Wang, Chi-Ping Day, Chih-Hsuan Wei, Qiao Jin, Robert Leaman, Yifan Yang, Shubo Tian, Aodong Qiu, Yin Fang, Qingqing Zhu, Xinghua Lu, Zhiyong Lu.

2024 **Demystifying Large Language Models for Medicine: A Primer.** Qiao Jin, Nicholas Wan, Robert Leaman, Shubo Tian, Zhizheng Wang, Yifan Yang, Zifeng Wang, Guangzhi Xiong, Po-Ting Lai, Qingqing Zhu, Benjamin Hou, Maame Sarfo-Gyamfi, Gongbo Zhang, Aidan Gilson, Balu Bhasuran, Zhe He, Aidong Zhang, Jimeng Sun, Chunhua Weng, Ronald M Summers, Qingyu Chen, Yifan Peng, Zhiyong Lu.

2023 **Improving Fairness in Medical Imaging Through Causal Learning.** Yifan Yang, Xiaoyu Liu, Mingquan Lin, Yifan Peng, Furong Huang, Zhiyong Lu.

Talks

2024

Adversarial Attack on Large Language Models in Medicine. *AMIA 2024*

2024 Fairness in biomedical AI. *University of Delaware*

Internships

Summer 2022 Intern - Data Science and Statistical Computing - Visualization and Interactive Data Analytics, gRED, Genentech.