Zihao(Gavin) Yang
New York, NY 10009 | 646-897-9271 | zy2091@nyu.edu

EDUCATION

New York University New York, NY
B.A. in Computer Science and Data Science, Minor in Mathematics *Sept 2020 – May 2024*

- Cumulative GPA: 3.82/4.0; Last 2 years GPA: 4.0/4.0
- Honors: George Maker Research Scholar, Dean's Honors List, Dean's Undergraduate Research Fund Grant, Cum Laude

PUBLICATION

- Yanbing Chen, Ruilin Wang, **Zihao Yang**, Lavender Jiang, Eric Oermann. Refining Packing and Shuffling Strategies for Enhanced Performance in Generative Language Models. *Under review at **ACL ARR***, 2024.
- Daniel Alber, **Zihao Yang**, Sumedha Rai, Eunice Yang, Aly Valliani, Gabriel Rosenbaum, Ashley Amend-Thomas, David Kurland, Monika Hedman, Caroline Kremer, Alexander Eremiev, Bruck Negash, Daniel Wiggan, Michelle Nakatsuka, Karl Sangwon, Sean Neifert, Hammad Khan, Akshay Save, Xujin Liu, Lavender Jiang, Daniel Orringer, Douglas Kondziolka, Eric Oermann. Medical large language models are vulnerable to attack. *Under Review at **Nature***, 2024.
- Chi Hang, Ruiqi Deng, Lavender Jiang, **Zihao Yang**, Daniel Alber, Anton Alyakin, Eric Oermann. BPQA Dataset: Evaluating How Well Language Models Leverage Blood Pressures to Answer Biomedical Questions. *Preprint*, 2024.
- Lavender Jiang, Daniel Alber, **Zihao Yang**, Karl Sangwon, Xujin Liu, Kyunghyun Cho, Eric K. Oermann. Language Models Can Guess Your Identities from De-identified Clinical Notes. *Preprint*, 2024.
- **Zihao Yang**, Chenkang Zhang, Muru Wu, Xujin Liu, Lavender Jiang, Kyunghyun Cho, Eric Oermann. Intriguing Effect of the Correlation Prior on ICD-9 Code Assignment. *Proceedings of the 61st Annual Meeting of the **Association for Computational Linguistics** (Volume 4: Student Research Workshop)*, 2023.

RESEARCH EXPERIENCE

Efficient Multimodal Language Model New York, NY
NYU VisionX Lab, Advisor: Dr. Saining Xie *July 2024-Present*

- Proposed to compress multimodal language models and improve their efficiency to enhance accessibility and broaden their applications.
- Migrated Cambrian codebase from TPU setup to GPU setup and trained Cambrian models based on small language models like Phi3-mini model.
- Explored vision token compression techniques for reducing computational cost, such as Token Merging and VoCo-LLaMA.

Training Dynamics of Medical LLMs New York, NY
NYU Langone OLAB, Advisor: Dr. Eric K. Oermann *Feb 2024 – May 2024*

- Explored several research questions related to different aspects of training dynamics with the goal of training a trustworthy medical large language model (LLM).
- Trained multiple 1-billion parameter models using different data mixtures consisting of Starcoder, PubMed abstracts, and Slimpajama, to study how pretraining data mixtures affect downstream performance.
- Studied the effect of data cleaning on both the performance of LLMs and their memorization of sensitive information.

Language Model Neurosurgical Benchmark New York, NY
NYU Langone OLAB, Advisor: Dr. Eric K. Oermann *Sept 2023 – July 2024*

- Created a multiple choice question answering benchmark that includes 2 datasets of specialized neurosurgical questions from board examinations.
- Compared the performance of various current LLMs on this benchmark with their performance on 5 widely

used datasets of general medical questions.

- Measured question difficulty based on the performance of LLMs and revealed the nature of difficult questions by using topic modeling.

Data Poisoning Attacks on Medical LLMs New York, NY

NYU Langone OLAB, Advisor: Dr. Eric K. Oermann *July 2023 – July 2024*

- Investigated how vulnerable LLMs are to injection of medical misinformation into the training dataset.
- Demonstrated the invisibility of data poisoning attacks and the resulting medical misinformation through evaluation on existing medical benchmarks.
- Proposed a method to validate LLM outputs using a verified biomedical knowledge graph.

Common Patterns in Spatial Memory New York, NY

NYU Langone Wisniewski Lab, Advisors: Dr. Thomas M. Wisniewski; Dr. Shuo Chen *Nov 2022 – Mar 2024*

- Explored mechanisms underlying spatial memory encoding and storage in hippocampus through analyzing electrophysiological signal recordings from mice.
- Employed a hybrid approach by integrating traditional signal processing techniques and machine learning methods to identify generalizable patterns associated with spatial memory.

Intriguing Effect of Correlation Prior on ICD-9 Code Assignment New York, NY

NYU Langone OLAB, Advisors: Dr. Eric K. Oermann; Dr. Kyunghyun Cho *June 2022 – May 2023*

- Investigated how incorporating correlation prior into language models affects their performance on predicting clinical diagnosis and procedure codes from clinical texts such as discharge summaries.
- Proposed training with two different kinds of clinical codes as a passive method to incorporate correlation prior, and used a regularization technique as an active method to incorporate correlation prior.
- Conducted ablation experiments to demonstrate the effects of methods used to incorporate correlation prior.

Perceiver in Long-range Language Context New York, NY

Advisor: Dr. Samuel R. Bowman *Feb 2022 – May 2022*

- Evaluated the reasoning capabilities of Perceiver on tasks involving long sequence texts using the Long Range Arena benchmark.
- Employed the vanilla transformer and several efficient transformers as baselines to conduct comparative research on Perceiver.

SKILL

Programming Language and Tool: Java, Python, R, MATLAB, SQL, Linux, Git

DS & ML Library: PyTorch, Hugging Face, DeepSpeed, Dask, WandB, Ray Tune

Language: Proficient in Chinese and English