Zichun **Yu**

6413 Gates Hillman Complex, 4902 Forbes Ave, Pittsburgh, PA 15213

zichunyu@andrew.cmu.edu | yuzc19.github.io | Google Scholar

Education

Carnegie Mellon University (CMU)

PhD student at Language Technologies Institute *Aug. 2023 - Present*

- GPA: 4.0/4.0
- Core Courses: Large Language Models Methods and Application (A+), Multimodal Machine Learning (A+), On-Device Machine Learning (A+), Introduction to Machine Learning (A), Algorithmic Foundations of Interactive Learning (A)

Tsinghua University (THU)

B.Eng. in Computer Science and Technology *Sep. 2019 - Jul. 2023*

- GPA: 3.94/4.0, Ranking: 7/202
- Core Courses: Computer Organization (100, Top 1), Operating Systems (100, Top 1), Formal Languages and Automata (100, Top 1), Probability and Statistics (100, Top 1), Foundation of Object-Oriented Programming (99, Top 3)

Research Interests

- ***Intelligent and efficient LLM scaling with novel pretraining data curation and synthesis methods.***
- ***Data valuation and influence attribution to better capture the impact of LLM training data.***

Selected Publications

- **RePro: Training Language Models to Faithfully Recycle the Web for Pretraining** (In Review) [PDF]
  Zichun Yu, Chenyan Xiong
  TL;DR: Training a small LM with carefully designed RL to faithfully rephrase web text into higher-quality pretraining data.
- **Group-Level Data Selection for Efficient Pretraining** (NeurIPS'25) [PDF]
  Zichun Yu, Fei Peng, Jie Lei, Arnold Overwijk, Wen-tau Yih, Chenyan Xiong
  TL;DR: Group-level data selection method that improves pretraining efficiency by modeling interactions among training data.
- **FLAME-MoE: A Transparent End-to-End Research Platform for Mixture-of-Experts Language Models** (arXiv) [PDF]
  Hao Kang*, **Zichun Yu***, Chenyan Xiong
  TL;DR: An open, end-to-end, cross-scale research platform for MoE models, complete with code, data, logs, and checkpoints.
- **Montessori-Instruct: Generate Influential Training Data Tailored for Student Learning** (ICLR'25) [PDF]
  Xiaochuan Li, **Zichun Yu**, Chenyan Xiong
  TL;DR: Training a teacher model to generate synthetic data tailored to a student model's learning through data influence.
- **MATES: Model-Aware Data Selection for Efficient Pretraining with Data Influence Models** (NeurIPS'24) [PDF]
  Zichun Yu, Spandan Das, Chenyan Xiong
  TL;DR: Model-aware data selection that trains a small data influence model to pick influential pretraining examples on the fly.
- **An In-depth Look at Gemini's Language Abilities** (arXiv) [PDF]
  Syeda Nahida Akter*, **Zichun Yu***, Aashiq Muhamed*, Tianyue Ou*, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, Graham Neubig
  TL;DR: Holistic and in-depth evaluation of Gemini's abilities across knowledge, reasoning, coding, and multilingual tasks.
- **Augmentation-Adapted Retriever Improves Generalization of Language Models as Generic Plug-In** (ACL'23) [PDF]
  Zichun Yu, Chenyan Xiong, Shi Yu, Zhiyuan Liu
  TL;DR: Adapting a retriever based on LM preferences on documents to improve LM zero-shot generalization on knowledge tasks.