Last Updated on Nov 15th, 2025

Jingyu Liu

jingyu6.github.io Jingyu6 jingyu6 jingyu6@uchicago.edu Google Scholar

Education

University of Chicago — Sept. 2024 – Present
PhD in Computer Science, Advisor: Ce Zhang — *Chicago, IL, USA*

ETH Zürich — Sept. 2021 – Aug. 2024
MS in Computer Science with Major in Machine Intelligence — *Zurich, Switzerland*

New York University — Aug. 2016 – May 2020
BA with Honors in Computer Science (Major GPA: 3.97/4.00, Overall: 3.88/4.00) — *New York, NY, USA*

Research Interests

- Efficient training and inference of large language models.
- Diffusion language models and hybrid architectures.
- Natural language processing (code generation and long-context models).

Publications & Preprints

[9] **TiDAR: Think in Diffusion, Talk in Autoregression**
Jingyu Liu*, Xin Dong*, Zhifan Ye, Rishabh Mehta, Yonggan Fu, Vartika Singh, Jan Kautz, Ce Zhang, Pavlo Molchanov (Nvidia)

[8] **Efficient-DLM: From Autoregressive to Diffusion Language Models and Beyond in Speed**
Yonggan Fu*, Lexington Whalen*, Zhifan Ye, Xin Dong, Shizhe Diao, **Jingyu Liu**, Chengyue Wu, Hao Zhang, Enze Xie, Song Han, Maksim Khadkevich, Jan Kautz, Yingyan Celine Lin, Pavlo Molchanov (Submitted to ICLR 2026)

[7] **HAMburger: Accelerating LLM Inference via Token Smashing**
Jingyu Liu, Ce Zhang

[6] **Speculative Prefill: Turbocharging TTFT with Lightweight and Training-Free Token Importance Estimation**
Jingyu Liu, Beidi Chen, Ce Zhang (ICML 2025)

[5] **How Far Are We From AGI? Are LLMs All We Need?**
Tao Feng*, Chuanyang Jin*, **Jingyu Liu***, Kunlun Zhu*, Haoqin Tu, Zirui Cheng, Guanyu Lin, Jiaxuan You (TMLR 2024)

[4] **Effective Long-Context Scaling of Foundation Models**
Wenhan Xiong*, **Jingyu Liu***, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oğuz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, Hao Ma (NAACL 2024)

[3] **Code Llama: Open Foundation Models for Code**
Baptiste Rozière*, Jonas Gehring*, Fabian Gloeckle*, Sten Sootla*, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, **Jingyu Liu**, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, Gabriel Synnaeve* (Meta AI)

[2] **CLIP-Layout: Style-Consistent Indoor Scene Synthesis with Semantic Furniture Embedding**
Jingyu Liu, Wenhan Xiong, Ian Jones, Yixin Nie, Anchit Gupta, Barlas Oğuz

[1] **Text-guided 3D Human Generation from 2D Collections**

Tsu-Jui Fu, Wenhan Xiong, Yixin Nie, **Jingyu Liu**, Barlas Oğuz, William Yang Wang (EMNLP 2023)

[0] **Scene-LLM: Extending Language Model for 3D Scene Reasoning**

Rao Fu, **Jingyu Liu**, Xilun Chen, Yixin Nie, Wenhan Xiong (WACV 2025)

Experience

Nvidia Jun. 2025 – Dec. 2025

Research Scientist Intern *Santa Clara, CA, USA*

- Building SOTA diffusion LLMs with hybrid AR and Diffusion architecture. Exploring other diffusion architectures such as Block Diffusions.
- Scaling foundational diffusion LLMs including MDLM, EsoLM, and DUO models with distillation techniques.

Together AI Jan. 2024 – Aug. 2024

Research Assistant at Llama *Turbo Team* *Remote*

- Worked on continual pretraining and instruction fine-tuning of foundation models.
- Explored grafting low-rank MLPs and Sliding-Window Attention (SWA) to pretrained Llama 70B and 405B models to reduce inference FLOPs.
- Engineered pretraining frameworks to support tensor parallelism and continuous batching during training. Developed comprehensive frameworks for model evaluation.

Meta Aug. 2022 – Aug. 2023

AI Resident at Gen AI *Menlo Park, CA, USA*

- Worked on the SOTA open-sourced code generation LLMs, CodeLlama [4], including context extension & robust and efficient programming problem evaluations on the family of models, ranging from interview to advanced difficulties.
- Worked on extending the context window of Llama 2. Our Llama 2 Long [3] beats gpt-3.5-16k on a wide range of long context benchmarks and Llama 2 on all evaluated tasks. Conducted both large scale pretraining and finetuning experiments and provided analysis on all essential design choices (data, architecture, sparse attention, etc) for effective context scaling.
- Research on semantic indoor scene synthesis from text prompts [2] using projected CLIP features and permutation-invariant Transformers. Worked on text-guided 3D human generation from 2D data [1], which adopts cross-modal attention to fuse compositional human rendering with the extracted fashion semantics. Worked on finetuning Llama 2 with extracted 3D features for semantic 3D scene understanding and reasoning [0].

ETH Zürich Mar. 2022 – Nov. 2022

Research Assistant, LAS Lab led by professor Andreas Krause *Zürich, Switzerland*

- The project aimed to design an offline reinforcement learning algorithm that can performance well when we are given a mixture dataset that consists of trajectories from multiple demonstrators. The goal is that the RL agent can achieve at least as good performance as if it is trained with the data from the best policy alone and even eclipse them when the extra sub-optimal data can provide "useful information" about the task.
- We proposed a variant of CQL algorithm called *Expert-Regularized CQL* (erCQL, code available at repo) that solves the problem in a restricted setting where there is a dominating policy and the source of each transition sample is known.
- Our erCQL first behavior-clones the best policy and uses it to relabel transitions from sub-optimal data sources. Then the agent is trained with the same objective except that the actions of all transitions are predicted with our BC policy.
- We showed that in many OpenAI gym tasks, erCQL outperforms CQL in almost all data mixtures and can often beat CQL trained with only data from the best policy in terms of convergence rate and final score.

ByteDance Aug. 2020 – Aug. 2021

Machine Learning Engineer *Beijing, China*

- Worked on the e-commerce search engine for Douyin, a platform for live-streamers to search products to sell.
- Built the ranking module from MVP stage to fully functional service, which includes the data processing pipeline with Kafka, training instance joining & feature extraction using using PySpark, model design, model training, automatic deployment and daily update on the company's own deep learning eco-system.

- Used Elastic Search for item retrieval and distributed LambdaMART for pre-ranking; Built services with BERT, Kernel-based Neural Ranking Model, and Electra for query parsing & understanding; Implemented Wide-and-Deep & variants of Deep-FM in TensorFlow for ranking (CTR, CVR prediction).
- Improved CTR from low baseline of 20% to over 50%. Iterated over multiple versions and led the development of new product features as well as the group discussion of SOTA works about search & recommender systems.
- Received highest rating from the team on both the technical achievement and communication in the annual performance review.

New York University Sept. 2018 – May. 2019
Computer System Organization Tutor *New York, USA*

- Tutor concepts such as cache memory, virtual memory, X86 assembly code, malloc library, data representation, and multi-threading.

Raycloud Technology June 2018 – Aug. 2018
Algorithm Engineer Intern *Hangzhou, China*

- Worked for Kuai Mai Design, a product for helping e-commerce sellers automatically generate information pages with deep learning and computer vision algorithms.
- Implemented Hungary bipartite matching algorithm to match user-provided images with available slots in PSD template based on pose similarity, color consistency, and image content type.

Side Projects

Conservative Offline Q-Learning with Gaussian Processes [*code*] [*report*] Oct. 2021 – Feb. 2022
Deep Q Learning with Backward SARSA [*code*] [*report*] Sept. 2021 – Feb. 2022
CycleGAN with Shape-Color Regularization [*code*] [*report*] Feb. 2019 – May 2019
Downpour Asynchronous Stochastic Gradient Descent Mar. 2019 – May 2019

Reviews

ICLR: 2025 (How Far Are We From AGI workshop), 2026
ICML: 2024 (Long-Context Foundation Models workshop), 2025

Honors and Awards

2020 Undergraduate Prize for Outstanding Performance in Computer Science, NYU
(Up to three recipients in the department per year)
2016-2020 College of Arts and Science Scholarship & Tisch School Scholarship, NYU
2016-2020 Dean's List, NYU

Skills

Programming Languages : Python, C++, C, Go, Java, C#, MySQL, LaTeX
Deep Learning Framework : PyTorch, vLLM, xFormers, FairScale, TensorFlow
Game Engines and Software : Linux, Unity3D, Office, Adobe Photoshop, Blender, Autodesk Maya