Hongye Jin

Homepage / Google Scholar

Applied Scientist, Amazon Palo Alto, CA mooler0410@gmail.com

Professional Experience

- **At Amazon Rufus Midtrain/Post-train Team**
  - Core contributor to SFT/RL post-training infra, making it work for our 100B and 800B ultra-sparse MoE models.
  - Built the team's post-training data decontamination pipeline.
- **At Amazon Rufus Pretrain Team**
  - Core member of the 100B/800B ultra-sparse MoE model pretraining effort. Drove key observations on early-layer MoE behavior and improved expert training dynamics.Mitigate Silent Expert Death in Ultra-Sparse MoE (https://alltoall.notion.site/save-lower-layer-moe-experts-llal)
  - Owned the long context extension recipe for our 400B MoE and 100B/800B ultra sparse MoE models. Identified effective training strategies and hyperparameters, and implemented the corresponding changes in our pretraining framework. Also Built long-context and sparse-attention scaling-law experimentation toolkits, including benchmarks, scaling-law configurations, and sparse-attention methodologies.
- **Before Amazon Rufus** (**Efficient LLMs**)
  - Developed Self-Extend (https://github.com/datamllab/LongLM), a pioneering training-free method for context-window extension. (ICML 2024 Spotlight; featured at Google I/O).

Experience

Amazon, Applied Scientist, Palo Alto, CA — March 2025 – Present
- Working on large scale LLM post-training, RL infra, and Base modeling

Amazon, Palo Alto, CA — May 2024 – Aug 2024
- *Applied Scientist Intern*, Large Language Model Base modeling

Visa Research, Palo Alto, CA — Sept 2022 – Dec 2022
- *Research Intern*, Graph Neural Networks

DAMO Academy, Alibaba, Beijing, China — Nov 2020 – Feb 2021
- *Research Assistant*, Distant Supervision for NLP

NExT++ Lab, NUS, Singapore — Sept 2019 – March 2020
- *Research Assistant*, GNNs for Recommender system

Publications [Google Scholar]

Selected recent work. The full list can be found on Google Scholar. 3,500+ citations.

[*arXiv*] **Stepwise Penalization for Length-Efficient Chain-of-Thought Reasoning**
Xintong Li, Sha Li, Rongmei Lin, **Hongye Jin**, Linwei Li, Hejie Cui, Sarah Zhang, Chia-Yuan Chang, Kewei Cheng, Besnik Fetahu, Priyanka Nigam, Jingbo Shang, Bing Yin
arXiv, 2026

[NeurIPS2025] **Longer Context, Deeper Thinking: Uncovering the Role of Long-Context Ability in Reasoning**
Wang Yang, Zirui Liu, **Hongye Jin**, Qingyu Yin, Vipin Chaudhary, Xiaotian Han
Neural Information Processing Systems (NeurIPS), 2025

[ACL2025] **💯-LongBench: Are de facto Long-Context Benchmarks Literally Evaluating Long-Context Ability?**
Wang Yang, **Hongye Jin**, Shaochen Zhong, Song Jiang, Qifan Wang, Vipin Chaudhary, Xiaotian Han
ACL Findings, 2025

[*arXiv*] **Thinking Preference Optimization**
Wang Yang, **Hongye Jin**, Jingfeng Yang, Vipin Chaudhary, Xiaotian Han
arXiv, 2025

[ICML2024] **LLM Maybe LongLM: Self-Extend LLM Context Window Without Tuning**
Hongye Jin[*], Xiaotian Han[*], Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, Xia Hu
International Conference on Machine Learning (ICML), 2024
ICML2024 Spotlight (3.5%)

[ICML2024] **Kivi: A tuning-free asymmetric 2bit quantization for kv cache**
Zirui Li[*], Jiayi Yuan[*], **Hongye Jin**, Shaochen (Henry) Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, Xia Hu
International Conference on Machine Learning (ICML), 2024

[TKDD2024] **Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond**
Jingfeng Yang[*], **Hongye Jin**[*], Ruixiang Tang[*], Xiaotian Han[*], Qizhang Feng[*], Haoming Jiang, Bing Yin, Xia Hu
Transactions on Knowledge Discovery from Data (TKDD), 2024
Featured a widely used tree-style taxonomy of LLM, my favorite Art work

[NeurIPS2023] **Chasing Fairness under Distribution Shift: a Model Weight Perturbation Approach**
Zhimeng Jiang[*], Xiaotian Han[*], **Hongye Jin**, Guanchu Wang, Rui Chen, Na Zou, Xia Hu
Neural Information Processing Systems (NeurIPS), 2023

[TMLR2023] **Retiring ΔDP: New Distribution-Level Metrics for Demographic Parity**
Xiaotian Han[*], Zhimeng Jiang[*], **Hongye Jin**[*], Zirui Liu, Na Zou, Qifan Wang, Xia Hu
Transactions on Machine Learning Research (TMLR), 2023

[*arXiv*] **GrowLength: Accelerating LLMs Pretraining by Progressively Growing Training Length**
Xiaotian Han[*], **Hongye Jin**[*], Jingfeng Yang, Zhimeng Jiang, Chia-Yuan Chang, Xia Hu
arXiv, 2023

Education

- *Ph.D. in Computer Science*, Texas A&M University, College Station, TX — Aug 2020 – May 2025
- *B.S. in Computer Science*, Peking University — Sep 2016 – June 2020
- *Chemistry*, Peking University — Sep 2015 – Aug 2016

Professional Services

Program Committee/Reviewer
ICLR 2024, 2025; ICML 2023-2025; NeurIPS 2023-2025; AAAI 2023-2025; IJCAI 2023; WWW 2024

Last updated on 08/23/2026