

# KAIYUE WEN

☎ 6502106367 / 13670156595

✉ [kaiyuewen3@gmail.com](mailto:kaiyuewen3@gmail.com)

📄 [Personal Site](#)

🌐 [github.com/WhenWen](https://github.com/WhenWen)

## Education

Phd, Computer Science Department, Stanford University

Expected 07/2029

BS, Institute for Interdisciplinary Information, Tsinghua University

09/2020-07/2024

Overall GPA: 3.95/4.00

## Awards and Honors

### Competitive Mathematics

1 <sup>st</sup> Prize in National High School Mathematics Olympics Competition	11/2019
1 <sup>st</sup> Prize in National High School Mathematics Olympics Competition	11/2018
Silver Medal in S.-T. Yau College Student Mathematics Contest on Probability and Statistics (rank 3)	05/2021
Bronze Medal in S.-T. Yau College Student Mathematics Contest Team Track	05/2021
Silver Medal in S.-T. Yau College Student Mathematics Contest on Probability and Statistics (rank 3)	09/2022
Silver Medal in S.-T. Yau College Student Mathematics Contest Team Track (rank 2)	09/2022

### Honors

Comprehensive Merit Scholarship of Tsinghua	10/2021
Comprehensive Merit Scholarship of Tsinghua	10/2022
Silver Medal in Yao Award (top scholarship in IIS; 3 students institute-wide)	09/2023
National Scholarship (top 0.2% national-wide)	10/2023
Stanford Graduate Fellowship	09/2024

## Publications

(\* stands for equal contribution.)

- [1](**ICLR 2026**) **Kaiyue Wen**, David Hall, Tengyu Ma, Percy Liang. “Fantastic Pretraining Optimizers and Where to Find Them”
- [2](**ICLR 2026**) Jiazhen Li, Hongzhou Lin, Hong Lu, **Kaiyue Wen**, Zaiwen Yang, Jiaxuan Gao, Yi Wu, Jingzhao Zhang. “QuestA: Expanding reasoning capacity in llms via question augmentation”
- [3](**NeurIPS 2025, Oral, Best Paper**) Zihan Qiu\*, Zekun Wang\*, Bo Zheng\*, Zeyu Huang\*, **Kaiyue Wen**, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, Junyang Lin. “Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free”
- [4](**NeurIPS 2025**) Songlin Yang, Yikang Shen, **Kaiyue Wen**, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, Yoon Kim. “PaTH Attention: Position Encoding via Accumulating Householder Transformations”
- [5](**ACL 2025**) Zihan Qiu\*, Zeyu Huang\*, Bo Zheng\*, **Kaiyue Wen**, Zekun Wang, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, Junyang Lin. “Demons in the Detail: On Implementing Load Balancing Loss for Training Specialized Mixture-of-Expert Models”
- [6](**COLM 2025**) Xingyu Dang, Christina Baek, **Kaiyue Wen**, Zico Kolter, Aditi Raghunathan. “Weight Ensembling Improves Reasoning in Language Models”
- [7](**ICML 2025**) Amirhesam Abedsoltan, Huaqing Zhang, **Kaiyue Wen**, Hongzhou Lin, Jingzhao Zhang, Mikhail Belkin. “Task Generalization With AutoRegressive Compositional Structure: Can Learning From D Tasks Generalize to  $D^T$  Tasks?”
- [8](**ICML 2025**) Jacob Mitchell Springer, Sachin Goyal, **Kaiyue Wen**, Tanishq Kumar, Xiang Yue, Sathika Malladi, Graham Neubig, Aditi Raghunathan. “Overtrained Language Models Are Harder to Fine-Tune”
- [9](**ICLR 2025**) **Kaiyue Wen**\*, Xingyu Dang\*, Kaifeng Lyu. “RNNs are not Transformers (Yet): The Key Bottleneck on In-context Retrieval”
- [10](**ICLR 2025**) **Kaiyue Wen**, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, Tengyu Ma. “Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective”
- [11](**ICLR 2025**) **Kaiyue Wen**\*, Huaqing Zhang\*, Hongzhou Lin, Jingzhao Zhang “From Sparse Dependence to Sparse Attention: Unveiling How Chain-of-Thought Enhances Transformer Sample Efficiency”
- [12](**Annals of Statistics, 2025**) **Kaiyue Wen**\*, Tengyao Wang\*, Yuhao Wang. “Residual Permutation Test for Regression Coefficient Testing”

- [13](**NeurIPS 2023, Oral**) **Kaiyue Wen**, Zhiyuan Li, Tengyu Ma. “Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization”
- [14](**NeurIPS 2023**) **Kaiyue Wen**, Yuchen Li, Bingbin Liu, Andrej Risteski. “(Un) interpretability of Transformers: a case study with Dyck grammars”
- [15](**ICLR 2023**) **Kaiyue Wen**, Tengyu Ma, Zhiyuan Li. “How Sharpness-Aware Minimization Minimizes Sharpness?”
- [16](**ICLR 2023**) **Kaiyue Wen**<sup>\*</sup>, Jiaye Teng<sup>\*</sup>, Jingzhao Zhang. “Benign Overfitting in Classification: Provably Counter Label Noise with Larger Models”
- [17](**EMNLP 2022**) Xiaozhi Wang<sup>\*</sup>, **Kaiyue Wen**<sup>\*</sup>, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, Juanzi Li. “Finding Skill Neurons in Pre-trained Transformer-based Language Models”
- [18](**NAACL 2022**) Yusheng Su<sup>\*</sup>, Xiaozhi Wang<sup>\*</sup>, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, **Kaiyue Wen**, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, Jie Zhou “On Transferability of Prompt Tuning for Natural Language Processing”

## Preprints and Manuscripts

---

- [19](**Preprint**) Arvind Mahankali, **Kaiyue Wen**, Tengyu Ma. “Divide-and-Conquer CoT: RL for Reducing Latency via Parallel Reasoning”
- [20](**Preprint**) Zihan Qiu<sup>\*</sup>, Zeyu Huang<sup>\*</sup>, **Kaiyue Wen**<sup>\*</sup>, Peng Jin<sup>\*</sup>, Bo Zheng, Yuxin Zhou, Haofeng Huang, Zekun Wang, Xiao Li, Huaqing Zhang, Yang Xu, Haoran Lian, Siqi Zhang, Rui Men, Jianwei Zhang, Ivan Titov, Dayiheng Liu, Jingren Zhou, Junyang Lin “A Unified View of Attention and Residual Sinks: Outlier-Driven Rescaling is Essential for Transformer Training”
- [21](**Manuscript**) Haozhe Jiang<sup>\*</sup>, **Kaiyue Wen**<sup>\*</sup>, Yilei Chen. “Practically Solving LPN in High Noise Regimes Faster Using Neural Networks”

## Skills

---

**Languages:** Familiar with Python and has written C++, R, Matlab, Bash

**Maths:** Familiar with optimization theory, mathematics analysis, measure theory, linear algebra, abstract algebra, probability theory, statistics, causal inference, and discrete mathematics

**Leadership:** Class monitor of Yao Class from 2021 to 2024, vice president of the IIIS Student Union from 2023 to 2024

**Miscellaneous:** Chinese debating (PB: rank 2 school-wide), science fiction novel writing