Kaiyue Wen

kaiyuewen3@gmail.com | Website | GitHub | Google Scholar

I study the science of deep learning. My work combines theory and experiments to understand deep learning and develop new algorithms. I am interested in using AI to accelerate understanding of deep learning.

Education

Stanford University, PhD student in Computer Science 09/2024–
Advisors: Tengyu Ma and Percy Liang. Stanford Graduate Fellowship (2024).
Tsinghua University, BS, Yao's class; GPA: 3.95/4.00 09/2020–07/2024

Selected Research

Optimization and Training Dynamics

Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective

First author; ICLR 2025

Pretraining loss exhibits a river valley landscape, which resembles a deep valley with a river at its bottom. This explains how training with a high learning rate can make more progress than training with a low learning rate, even when its intermediate losses are higher.

Fantastic Pretraining Optimizers and Where to Find Them

First author on both papers; Part I: ICLR 2026; Part II: Hyperball Optimization (2026 preprint)

Rigorous optimizer benchmarking shows that vanilla Muon's efficiency advantage over AdamW shrinks with overtraining under constant weight decay. The follow-up work introduces Hyperball, which uses weight-normalized training to improve Muon's efficiency.

Reasoning and Self-Improvement

Trust the Critic More

Co-first author with Luke Bailey; 2026 preprint

A learned critic can guide reasoning RL using partial solutions. On Qwen3-4B trained on FineProofs-RL, AC2 exceeds GRPO's best mean score on IMO-ProofBench with approximately 60% fewer decoding FLOPs.

Priced Guidance: Can Language Models Generate Future Research Ideas?

First author; 2026 preprint

Priced Guidance uses "20 Questions" to measure the amount of information an LLM needs to recover a future research idea. This yields a lower bound on the probability of generating that idea without hints.

Publications and Accepted Papers

* Equal contribution. Workshop papers are explicitly labeled.

[1] **(COLM 2026)** Luke Bailey, **Kaiyue Wen**, Kefan Dong, Tatsunori Hashimoto, Tengyu Ma. "Scaling Self-Play with Self-Guidance"

[2] **(COLM 2026, Oral)** Zihan Qiu*, Zeyu Huang*, **Kaiyue Wen***, Peng Jin*, 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"

[3] **(ICLR 2026) Kaiyue Wen**, David Hall, Tengyu Ma, Percy Liang. "Fantastic Pretraining Optimizers and Where to Find Them"

[4] **(ICLR 2026)** Jiazheng Li, Hongzhou Lin, Hong Lu, **Kaiyue Wen**, Zaiwen Yang, Jiaxuan Gao, Yi Wu, Jingzhao Zhang. "QuestA: Expanding Reasoning Capacity in LLMs via Question Augmentation"

[5] **(HiLD 2026 Workshop)** Miao Lu, Zeyu Bian, **Kaiyue Wen**, Beining Wu, Siyu Chen, Tianhao Wang, Zhiyuan Li. "Towards Understanding Momentum Acceleration in River-Valley Loss Landscape"

[6] **(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"

[7] **(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"

[8] **(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"

[9] **(ACL 2025)** Shengguang Wu, Fan-Yun Sun, **Kaiyue Wen**, Nick Haber. "Symmetrical Visual Contrastive Optimization: Aligning Vision-Language Models with Minimal Contrastive Images"

[10] **(COLM 2025)** Xingyu Dang, Christina Baek, **Kaiyue Wen**, Zico Kolter, Aditi Raghunathan. "Weight Ensembling Improves Reasoning in Language Models"

[11] **(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?"

[12] **(ICML 2025)** Jacob Mitchell Springer, Sachin Goyal, **Kaiyue Wen**, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, Aditi Raghunathan. "Overtrained Language Models Are Harder to Fine-Tune"

[13] **(ICLR 2025) Kaiyue Wen***, Xingyu Dang*, Kaifeng Lyu. "RNNs are not Transformers (Yet): The Key Bottleneck on In-context Retrieval"

[14] **(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"

[15] **(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"

[16] **(Annals of Statistics, 2025) Kaiyue Wen***, Tengyao Wang*, Yuhao Wang. "Residual Permutation Test for Regression Coefficient Testing"

[17] **(NeurIPS 2023, Oral) Kaiyue Wen**, Zhiyuan Li, Tengyu Ma. "Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization"

[18] **(NeurIPS 2023) Kaiyue Wen**, Yuchen Li, Bingbin Liu, Andrej Risteski. "Transformers are uninterpretable with myopic methods: a case study with bounded Dyck grammars"

[19] **(ICLR 2023) Kaiyue Wen**, Tengyu Ma, Zhiyuan Li. "How Sharpness-Aware Minimization Minimizes Sharpness?"

[20] **(ICLR 2023) Kaiyue Wen***, Jiaye Teng*, Jingzhao Zhang. "Benign Overfitting in Classification: Provably Counter Label Noise with Larger Models"

[21] **(EMNLP 2022)** Xiaozhi Wang*, **Kaiyue Wen***, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, Juanzi Li. "Finding Skill Neurons in Pre-trained Transformer-based Language Models"

[22] **(NAACL 2022)** Yusheng Su*, Xiaozhi Wang*, 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

[23] **(Preprint, 2026)** Xingyu Dang*, **Kaiyue Wen***, Sadhika Malladi. "The Best Optimizer Depends on Batch Size"

[24] **(Preprint, 2026) Kaiyue Wen**, Tengyu Ma, Percy Liang. "Priced Guidance: Can Language Models Generate Future Research Ideas?"

[25] **(Preprint, 2026)** Haodong Wen*, **Kaiyue Wen***, Jiaye Teng. "Tight Transition Time Bounds for Separable Logistic Regression at the Edge of Stability"

[26] **(Preprint, 2026) Kaiyue Wen***, Luke Bailey*, Arvind Mahankali, Tengyu Ma. "Trust the Critic More"

[27] **(Preprint, 2026) Kaiyue Wen**, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, Percy Liang. "Fantastic Pretraining Optimizers and Where to Find Them II: Hyperball Optimization"

[28] **(Preprint, 2026)** Slim Barkallah*, Luke Bailey*, **Kaiyue Wen***, Mohammed Abouzaid, Tengyu Ma. "Pseudo-Formalization for Automatic Proof Verification"

[29] **(Preprint, 2026)** Huaqing Zhang, **Kaiyue Wen**, Tengyu Ma. "Configuration-to-Performance Scaling Law with Neural Ansatz"

[30] **(Preprint, 2026)** Arvind Mahankali, **Kaiyue Wen**, Tengyu Ma. "Divide-and-Conquer CoT: RL for Reducing Latency via Parallel Reasoning"

[31] **(Preprint, 2023)** Haozhe Jiang*, **Kaiyue Wen***, Yilei Chen. "Practically Solving LPN in High Noise Regimes Faster Using Neural Networks"