Shaobo Wang

gszfwsb@gmail.com · https://gszfwsb.github.io · Google Scholar · Linkedin

Research Interests

I study *scalable, efficient, and theoretically grounded* methods to ***optimize the data-centric lifecycle (training, inference, and evaluation) of foundation models***, where quality and efficiency must be jointly optimized.

Education Background

Shanghai Jiao Tong University, Doctor of Artificial Intelligence 2024 – 2028 (expected)
Advisor: Prof. Linfeng Zhang

Shanghai Jiao Tong University, Master of Computer Science and Technology 2022 – 2024
Advisor: Prof. Junchi Yan, GPA: 3.98/4.0, Rank: 2/81

Harbin Institute of Technology, Bachelor of Software Engineering 2018 – 2022
GPA: 3.96/4.0, Rank: 1/181

Research Internship

Qwen Team, Alibaba Group 2025.05 – Present
Research Intern on LLM pre-training, supervised by Dr. Dayiheng Liu and Xingzhang Ren.
My core responsibilities include (i) optimizing the pre-training data mixture and curation for the Qwen3.5 series, and (ii) conducting pioneering research into next-generation pre- and mid-training methodologies.

Representative Papers

* denotes equal contribution. † denotes corresponding author.

OPUS: Towards Efficient and Principled Data Selection in Large Language Model Pre-training in Every Iteration [1]
*arXiv preprint, **co-worked with Qwen Team***
Shaobo Wang, Xuan Ouyang, Tianyi Xu, Yuzheng Hu, Jialin Liu, Guo Chen, Tianyu Zhang, Junhao Zheng, Kexin Yang, Xingzhang Ren, Dayiheng Liu, Linfeng Zhang.

Dataset Distillation with Neural Characteristic Function: A Minmax Perspective [2]
*CVPR 2025 **Highlight, Full Marks, talk invited at BAAI***
Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, Linfeng Zhang.

Rethinking LLM Evaluation: Can We Evaluate LLMs with 200× Less Data? [3]
*ICLR 2026, **co-worked with Qwen Team & GLM Team***
Shaobo Wang, Cong Wang, Wenjie Fu, Yue Min, Mingquan Feng, Isabel Guan, Xuming Hu, Conghui He, Cunxiang Wang, Kexin Yang, Xingzhang Ren, Fei Huang, Dayiheng Liu, Linfeng Zhang

Socratic-Zero: Bootstrapping Reasoning via Data-Free Agent Co-evolution [4]
arXiv preprint
Shaobo Wang, Zhengbo Jiao, Zifan Zhang, Yilang Peng, Xu Ze, Boyu Yang, Wei Wang, Hu Wei, Linfeng Zhang.

Winning the Pruning Gamble: A Unified Approach to Joint Sample and Token Pruning for Efficient Supervised Fine-Tuning [5]
*arXiv preprint, **co-worked with Qwen Team***
Shaobo Wang, Jiaming Wang, Jiajun Zhang, Cong Wang, Yue Min, Zichen Wen, Fei Huang, Huiqiang Jiang, Junyang Lin, Dayiheng Liu, Linfeng Zhang.

Other Papers

Grounding and Enhancing Informativeness and Utility in Dataset Distillation [6]

ICLR 2026

Shaobo Wang†, Yantai Yang, Guo Chen, Peiru Li, Kaixin Li, Yufa Zhou, Zhaorun Chen, Linfeng Zhang

Gnothi Seauton: Empowering Faithful Self-Interpretability in Black-Box Models [7]

ICLR 2025

Shaobo Wang, Hongxuan Tang, Mingyang Wang, Hongrui Zhang, Xuyang Liu, Xuming Hu, Linfeng Zhang.

Data Whisperer: Efficient Data Selection for Task-Specific LLM Fine-Tuning via Few-Shot In-Context Learning [8]

ACL Main 2025

Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, Linfeng Zhang.

ImagebindDC: Compressing Multimodal Data with Imagebind-based Condensation [9]

*AAAI 2026, **co-worked with Bosch Corporate Research Asia Pacific***

Yue Min*, Shaobo Wang*, Jiaze Li, Tianle Niu, Junxin Fan, Yongliang Miao, Lijin Yang, Linfeng Zhang.

UNSEEN: Incremental Dataset Pruning via Cross-Model Generalization Scoring [10]

AAAI 2026

Furui Xu*, Shaobo Wang*, Jiajun Zhang, Chenghao Sun, Haixiang Tang, Linfeng Zhang.

DRUPI: Dataset Reduction Using Privileged Information [11]

ICLR 2025 Workshop

Shaobo Wang, Yantai Yang, Shuaiyu Zhang, Xuming Hu, Linfeng Zhang.

Not All Samples should be Utilized Equally: Towards Understanding and Improving Dataset Distillation [12]

CVPR 2025 Workshop

Shaobo Wang, Yantai Yang, Qilong Wang, Kaixin Li, Linfeng Zhang, Junchi Yan.

CircuitSeer: Mining High-Quality Data by Probing Mathematical Reasoning Circuits in LLMs [13]

arXiv preprint

Shaobo Wang†, Yongliang Miao, Yuancheng Liu, Qianli Ma, Ning Liao, Linfeng Zhang.

VideoCompressa: Data-Efficient Video Understanding via Joint Temporal Compression and Spatial Reconstruction [14]

arXiv preprint

Shaobo Wang†, Tianle Niu, Runkang Yang, Deshan Liu, Xu He, Zichen Wen, Conghui He, Xuming Hu, Linfeng Zhang.

Shifting AI Efficiency From Model-Centric to Data-Centric Compression [15]

arXiv preprint

Xuyang Liu*, Zichen Wen*, Shaobo Wang*, Junjie Chen, Zhishan Tao, Yubo Wang, Tailai Chen, Xiangqi Jin, Chang Zou, Yiyu Wang, Chenfei Liao, Xu Zheng, Honggang Chen, Weijia Li, Xuming Hu, Conghui He, Linfeng Zhang

Efficient Multi-modal Large Language Models via Progressive Consistency Distillation [16]

NeurIPS 2025

Zichen Wen, Shaobo Wang, Yufa Zhou, Junyuan Zhang, Qintong Zhang, Yifeng Gao, Zhaorun Chen, Bin Wang, Weijia Li, Conghui He, Linfeng Zhang.

Stop Looking for Important Tokens in Multimodal Language Models: Duplication Matters More [17]

EMNLP Main 2025

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, Linfeng Zhang.

Think2Drive: Brick by Brick to Build a Model-based RL Agent for Quasi-Realistic Autonomous Driving (in CARLA-v2) [18]

ECCV 2024

Qifeng Li, Xiaosong Jia, Shaobo Wang, Junchi Yan.

Visualizing the Emergence of Intermediate Visual Patterns in DNNs [19]

NeurIPS 2021

Mingjie Li, Shaobo Wang, Quanshi Zhang.

SpeCa: Accelerating Diffusion Transformers with Speculative Feature Caching [20]

ACMMM 2025

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Fei Ren, Shaobo Wang, Kaixin Li, Linfeng Zhang.

Compute Only 16 Tokens in One Timestep: Accelerating Diffusion Transformers with Cluster-Driven Feature Caching [21]

ACMMM 2025

Zhixin Zheng, Xinyu Wang, Chang Zou, Shaobo Wang, Linfeng Zhang.

Visualizing the Emergence of Intermediate Visual Patterns in DNNs [22]

NeurIPS 2021

Mingjie Li, Shaobo Wang, Quanshi Zhang.

Reasoning Like an Economist: Post-Training on Economic Problems Induces Strategic Generalization in LLMs [23]

arXiv preprint

Yufa Zhou, Shaobo Wang, Xingyu Dong, Xiangqi Jin, Yifang Chen, Yue Min, Kexin Yang, Xingzhang Ren, Dayiheng Liu, Linfeng Zhang

dllm-cache: Accelerating diffusion large language models with adaptive caching [24]

arXiv, preprint

Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyuan Wei, Shaobo Wang, Linfeng Zhang.

A Survey of Linear Attention: Algorithm, Theory, Application, and Infrastructure [25]

TechRxiv preprint, ***co-worked with Tencent Hunyuan***

Yudong Zhang, Weixuan Sun, Xingwu Sun, Wei Ding, Ruiqi Xie, Hongyi Wang, Junjie Chen, Jiacheng Liu, Shaobo Wang, Yuwei Zhang, Yiqing Huang, Jiaming Wang, Tianchen Zhao, Weidong Han, Yanfeng Chen, Kai Zhang, Shuaipeng Li, Ruobing Xie, Di Wang, Jiansheng Chen, Linfeng Zhang, Chengzhong Xu, Yu Wang

DD-Ranking: Rethinking the Evaluation of Dataset Distillation [26]

Zekai Li, Xinhao Zhong, Samir Khaki, Zhiyuan Liang, Yuhao Zhou, Mingjia Shi, Ziqiao Wang, Xuanlei Zhao, Wangbo Zhao, Ziheng Qin, Mengxuan Wu, Pengfei Zhou, Haonan Wang, David Junhao Zhang, Jia-Wei Liu, Shaobo Wang, et al.

HONORS & AWARDS

Tencent Hunyuan Scholar, Tencent Hunyuan (one of 23 recipients in China) 2025