Qian Liu(刘乾)

✉ liuqian.sea@gmail.com · SivilTaram · https://siviltaram.github.io · SivilTaram

Experience

TikTok, Singapore — 2024.11 – Present
Researcher, *TikTok AI Innovation Center*

Sea AI Lab, Singapore — 2022.7 – 2024.11
Research Scientist & Team Lead, *Natural Language Processing Group*

Microsoft Research Asia, China — 2018.2 – 2022.6
Research Intern, *Data, Knowledge and Intelligence Group* ♢ Supervisor: Jian-Guang Lou & Bei Chen

Microsoft Research Asia, China — 2016.7 – 2017.8
Research Intern, *Big Data Mining Group* ♢ Supervisor: Zaiqing Nie & Yohn Cao

Education

Beihang University — 2017.9 – 2022.6
Ph.D. in Computer Science and Engineering ♣ Joint Program with Microsoft Research Asia
Ph.D. Supervisor: Qinping Zhao

Beihang University — 2013.9 – 2017.6
B.S. in Computer Science and Technology, Ranking 7 / 233

Research Interests

- **Code Generation**: *Building responsible, reliable and interpretable code generation models.*
- **Language Reasoning**: *Improving fundamental reasoning capabilities of language models.*

Selected Projects

Sailor2: Sailing in South-East Asia with Inclusive Multilingual LLMs — 2024.5 – 2024.12

We introduce Sailor2, a community-driven initiative that brings cutting-edge multilingual language models to South-East Asia (SEA). Our research highlights a strong demand for models in the 8B and 20B parameter range for production use, alongside a 1B model for specialized applications, such as speculative decoding and research purposes. These models, released under the Apache 2.0 license, provide enhanced accessibility to advanced language technologies across the region. As the **project leader**, I spearheaded project planning, team collaboration, and critical path identification. I conducted experiments on data cleaning, data deduplication, data mixture and secured computational resources while fostering active community engagement through discussions.

Sailor: Open Language Models for South-East Asia — 2023.4 – 2024.4

We present Sailor, a family of open language models ranging from 0.5B to 14B parameters, tailored for South-East Asian (SEA) languages. These models are continually pre-trained from Qwen1.5, a great language model for multilingual use cases. From Qwen1.5, Sailor models accept 200B to 400B tokens, primarily covering the languages of English, Chinese, Vietnamese, Thai, Indonesian, Malay, and Lao. The training leverages several techniques, including BPE dropout for improving the model robustness, aggressive data cleaning and deduplication, and small proxy models to optimize data mixture. As the **project leader**, I made comprehensive plans, implemented effective team collaboration, identified critical paths, and explored a variety of innovative technologies to drive the project forward.

Selected Publications

- Fangyu Lei*, Jixuan Chen*, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, Victor Zhong, Caiming Xiong, Ruoxi Sun, **Qian Liu**, Sida Wang, Tao Yu, Spider 2.0: Evaluating Language Models on Real-World Enterprise Text-to-SQL Workflows. (Preprint-2024)
- Siming Huang*, Tianhao Cheng*, J.K. Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J. Yang, J.H. Liu, Chenchen Zhang, Linzheng Chai, Ruifeng Yuan, Zhaoxiang Zhang, Jie Fu, **Qian Liu**, Ge Zhang, Zili Wang, Yuan Qi, Yinghui Xu, Wei Chu, OpenCoder: The Open Cookbook for Top-Tier Code Large Language Models. (Preprint-2024)
- Fan Zhou*, Zengzhi Wang*, **Qian Liu**, Junlong Li, Pengfei Liu, Programming Every Example: Lifting Pre-training Data Quality like Experts at Scale. (Preprint-2024)
- Chaofan Tao, **Qian Liu**✉, Longxu Dou✉, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, Ngai Wong, Scaling Laws with Vocabulary: Larger Models Deserve Larger Vocabularies. (Preprint-2024)
- **Qian Liu***, Xiaosen Zheng*, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, Min Lin, RegMix: Data Mixture as Regression for Language Model Pre-training. (Preprint-2024, * = equal contribution)
- BigCode, BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions. (Preprint-2024)
- Chengsong Huang*, **Qian Liu***, Bill Yuchen Lin*, Tianyu Pang, Chao Du, Min Lin, LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition. In *First Conference on Language Modeling* 2024 (COLM-2024, * = equal contribution)
- Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu✉, **Qian Liu**✉, Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (ACL-2024)
- Tongyao Zhu, **Qian Liu**✉, Liang Pang✉, Zhengbao Jiang, Min-Yen Kan, Min Lin, Beyond Memorization: The Challenge of Random Memory Access in Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (ACL-2024)
- Longxu Dou*, **Qian Liu***, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, Min Lin, Sailor: Open Language Models for South-East Asia. (Homepage, Preprint-2024, * = equal contribution)
- BigCode, StarCoder 2 and The Stack v2: The Next Generation. (Preprint-2024)
- Niklas Muennighoff, **Qian Liu**, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, Shayne Longpre, OctoPack: Instruction Tuning Code Large Language Models. In *International Conference on Learning Representations* 2024 (ICLR-2024 **Spotlight**)
- Yiheng Xu*, Hongjin Su*, Chen Xing*, Boyu Mi, **Qian Liu**, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, Zhoujun Cheng, Siheng Zhao, Lingpeng Kong, Bailin Wang, Caiming Xiong, Tao Yu, Lemur: Harmonizing Natural Language and Code for Language Agents. In *International Conference on Learning Representations* 2024 (ICLR-2024 **Spotlight**, * = equal contribution)
- Weichen Yu, Tianyu Pang✉, **Qian Liu**✉, Chao Du, Bingyi Kang, Yan Huang, Min Lin, Shuicheng Yan, Bag of Tricks for Training Data Extraction from Language Models. In *Proceedings of International Conference on Machine Learning* 2023 (ICML-2023)
- Tianping Zhang, Shaowen Wang, Shuicheng Yan, Jian Li, **Qian Liu**✉, Generative Table Pre-training Empowers Models for Tabular Prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 2023 (EMNLP-2023)
- BigCode, StarCoder: May the Source be with You! In *Transactions on Machine Learning Research* 2023 (TMLR 2023, **2.3K** Like in HuggingFace Hub)
- Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, **Qian Liu**, Rui Li, Xing Xie, Jian Tang, Learning on Large-scale Text-attributed Graphs via Variational Inference. In *International Conference on Learning Representations* 2023 (ICLR-2023 **Oral**)
- BigCode, SantaCoder: Don't Reach for the Stars! In *Deep Learning for Code Workshop on International Conference on Learning Representations* 2023 (DL4C@ICLR-2023 **Best Paper**)
- Xinyu Pi*, **Qian Liu***, Bei Chen, Morteza Ziyadi, Zeqi Lin, Yan Gao, Qiang Fu, Jian-Guang Lou, Weizhu

Chen, Reasoning Like Program Executors. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP-2022 **Oral**, MLADS-2022 **Distinguished Contribution Award**, * = equal contribution)

- **Qian Liu**, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, Jian-Guang Lou, TAPEX: Table Pre-training via Learning a Neural SQL Executor. In *International Conference on Learning Representations* 2022 (ICLR-2022 **Highest Rating in the 1st Round**)
- Jiaqi Guo, Ziliang Si, Yu Wang, **Qian Liu**, Ming Fan, Jian-Guang Lou, Zijiang Yang, Ting Liu, CHASE: A Large-Scale and Pragmatic Chinese Dataset for Cross-Database Context-Dependent Text-to-SQL. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (ACL-2021 **Oral**)
- Shuang Chen*, **Qian Liu***, Zhiwei Yu*, Chin-Yew Lin, Jian-Guang Lou, Feng Jiang, ReTraCk: A Flexible and Efficient Framework for Knowledge Base Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (ACL-2021 Demo, * = equal contribution)
- **Qian Liu***, Shengnan An*, Jian-Guang Lou, Bei Chen, Zeqi Lin, Yan Gao, Bin Zhou, Nanning Zheng, Dongmei Zhang, Compositional Generalization by Learning Analytical Expressions. In *Advances in Neural Information Processing Systems 34* (NeurIPS-2020 **Spotlight**, * = equal contribution)
- **Qian Liu**, Bei Chen, Jian-Guang Lou, Bin Zhou and Dongmei Zhang, Incomplete Utterance Rewriting as Semantic Segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP-2020)
- **Qian Liu**, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, Dongmei Zhang, You Impress Me: Dialogue Generation via Mutual Persona Perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL-2020)

♡ Service & Awards

Area Chair at ICLR 2025	2024
Co-Organizer of Deep Learning for The Third Table Representation Learning Workshop	2024
Area Chair at ACL 2024	2024
KAUST Rising Stars in AI 2024 (30 Worldwide)	2024
Beijing Outstanding Doctoral Thesis Nomination Awards	2023
Baidu Scholarship Nomination 2020 (20 Worldwide)	2020
Graduate Student National Scholarship	2019, 2021