Shimao Zhang

smzhang@smail.nju.edu.cn +86 150-3996-8336 https://shimao-zhang.github.io

Research Interests

My current research interests include Large Language Models, Reasoning, Interpretability, and Trustworthy AI. I'm also interested in expanding my research to multimodal scenarios in the future, including the physical world.

Education

Nanjing University **2023.09 – 2026.06**
Master in Computer Technology

- Co-advisors: Prof. Shujian Huang and Prof. Jiajun Chen
- GPA: 90.3/100

Nanjing University **2019.09 – 2023.06**
Bachelor in Computer Science and Technology

- Advisors: Prof. Shujian Huang and Prof. Lijun Zhang
- GPA: 86.0/100

Publications

[1] How does Alignment Enhance LLMs' Multilingual Capabilities? A Language Neurons Perspective
Shimao Zhang*, Zhejian Lai*, Xiang Liu*, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, Jiajun Chen. *In Proceedings of the 40th AAAI Conference on Artificial Intelligence.* **(AAAI'26 Oral)** [Link]

[2] Process-based Self-Rewarding Language Models
Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, Yeyun Gong. *In Findings of the Association for Computational Linguistics: ACL 2025.* **(ACL'25 Findings)** [Link]

[3] Getting More from Less: Large Language Models are Good Spontaneous Multilingual Learners
Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, Shujian Huang. *In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* **(EMNLP'24 Oral)** [Link]

[4] Distributed Projection-free Online Learning for Smooth and Convex Losses
Yibo Wang*, Yuanyu Wan*, **Shimao Zhang**, Lijun Zhang. *In Proceedings of the 37th AAAI Conference on Artificial Intelligence.* **(AAAI'23 Oral)** [Link]

[5] PATS: Process-Level Adaptive Thinking Mode Switching
Yi Wang*, Junxiao Liu*, **Shimao Zhang***, Jiajun Chen, Shujian Huang. *Preprint.* [Link]

[6] EDT: Improving Large Language Models' Generation by Entropy-based Dynamic Temperature Sampling
Shimao Zhang, Yu Bao, Shujian Huang. *Preprint.* [Link]

Internship Experience

Meituan *Beijing, CN*
Research Intern, M17 Foundation Model Group **2025.08 – Now**

- Mentor: Dr. Jiahuan Li
- Investigated the algorithms for foundation model optimization and LLMs mid-training.

Ant Group *Beijing, CN*
Research Intern, Ant Research **2025.07 – 2025.08**

- Mentor: Dr. Jian Guan
- Investigated the algorithms for VLMs, particularly long video understanding tasks.

Microsoft Research Asia (MSRA) *Beijing, CN*
Research Intern **2024.12 – 2025.06**

- Mentor: Dr. Xiao Liu and Dr. Yeyun Gong
- Conducted research on LLM reasoning and pretraining. Cleaned multilingual data and conducted research on data mixture for large-scale foundation model pretraining.

Research Experience

Nanjing University (NLP Group)

- **Co-advisors:** Prof. Shujian Huang and Prof. Jiajun Chen
- **Background:** Large Language Models have demonstrated outstanding performance across various downstream tasks and have been widely applied in multiple scenarios. This has prompted the urgent need for further enhancement and understanding of LLMs.
- **LLMs Reasoning & Generation:** We conduct research on LLM reasoning and generation by optimizing model training and generation strategies. We propose a novel paradigm called process-based self-rewarding language models [2], which effectively enhances LLMs' performance on complex mathematical reasoning tasks through an iterative self-rewarding pipeline. For generation strategies, we propose a dynamic entropy-based temperature sampling algorithm for a better trade-off between diversity and quality [6]. For generative long-thought reasoning, we propose a process-level adaptive thinking mode switching paradigm to dynamically switch the System 1/2 reasoning intensity for different reasoning steps [5].
- **Mechanistic Interpretability of LLMs:** To better understand the LLMs' mechanisms, we conduct a series of interpretability studies, particularly regarding multilingualism in LLMs. We first systematically reveal and analyze the spontaneous multilingual alignment phenomenon of LLMs [3], i.e., the model can achieve significant improvement even on languages unseen in the alignment. Furthermore, we propose a new language neuron identification algorithm [1], which is able to identify language-specific, language-related, and language-agnostic neurons in LLMs. Then we present empirical results and valuable insights that contribute to a deeper understanding of multilingual alignment and the multilingual capabilities of LLMs.

Northwestern University (MLL Lab)

- **Advisor:** Prof. Manling Li
- **Background:** LLMs often memorize sensitive, copyrighted, or harmful content from their vast training data, raising privacy, safety, and legal concerns. Thus, it is important to erase the specific knowledge via efficient post-training techniques.
- **Trustworthy and Responsible AI:** Machine unlearning aims to eliminate the impact of specific training data on the model. However, research on the side effects of existing algorithms is still very limited. We conduct a comprehensive investigation across multiple scenarios and methods. Additionally, we also provide valuable insights into the mechanistic interpretability and the side effects mitigation.

Nanjing University (LAMDA Group)

- **Advisor:** Prof. Lijun Zhang
- **Background:** Distributed online convex optimization has been a popular topic due to its powerful capability in online decision making, in which the projection operation could be the computational bottleneck.
- **Distributed Online Learning:** To avoid projections, distributed online projection-free methods have been proposed. However, they cannot utilize the smoothness condition, which has been exploited in the centralized setting to improve the regret. We propose a new distributed online projection-free method with a tighter regret for smooth and convex losses in the distributed setting [4].

Honours and Awards

- First-class Academic Scholarship, Nanjing University, 2023, 2024, 2025.
- Huatai Securities Technology Scholarship, Nanjing University, 2025
- Outstanding Graduate Student, Nanjing University, 2024.
- BYD Scholarship, Nanjing University, 2024.
- People's Scholarship, Nanjing University, 2021, 2022.
- First Prize, The 34th National High School Mathematics League, Chinese Mathematical Society, 2018.

Skills

- Programming: Python, C/C++, LaTeX, Git.
- Library: Pytorch, Transformers, Deepspeed, vLLM.
- Languages: Chinese, English (IELTS 7.0).
- Interests: Movie, Music, Football, Bodybuilding.