

Yiran Zhao

+65-91362816 | zhaoyiran@u.nus.edu | [Google Scholar](#) | [LinkedIn](#) | [Website](#)

EDUCATION

National University of Singapore

Doctor of Philosophy in Computer Science, Advised by Prof. Kenji Kawaguchi

Singapore

Aug. 2021 – now

Shanghai Jiaotong University

Bachelor of Engineering in Computer Science and Technology, GPA: 3.7/4.0

Shanghai, China

Sept. 2017 – Jul. 2021

- Zhiyuan Honours Bachelor Degree (**top 1%**)












RESEARCH INTERESTS

Interpretability, Multilingual, Alignment, and Reasoning within Large Language Models



OPEN SOURCE PROJECT (* DENOTES CO-FIRST AUTHORS)

- SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages W. Zhang*, HP. Chan*, **Y. Zhao***, M. Aljunied*, J. Wang*, C. Liu, Y. Deng, Z. Hu, W. Xu, YK. Chia, X. Li, L. Bing. (Responsible for training base models) [Website](#), [Model](#), [DEMO](#), [Github](#), [Technical Report](#)

PUBLICATIONS (* DENOTES CO-FIRST AUTHORS)

- Accelerating Greedy Coordinate Gradient via Probe Sampling  
Y. Zhao, W. Zheng, T. Cai, XL. Do, K. Kawaguchi, A. Goyal, M. Shieh, *accepted by NeurIPS 2024*
- How do Large Language Models Handle Multilingualism? 
Y. Zhao, W. Zhang, G. Chen, K. Kawaguchi, L. Bing, *accepted by NeurIPS 2024*
- Reasoning Robustness of LLMs to Adversarial Typographical Errors
E. Gan*, **Y. Zhao***, L. Cheng, Y. Mao, A. Goyal, K. Kawaguchi, MY. Kan, M. Shieh, *accepted by EMNLP 2024*
- Prompt Optimization via Adversarial In-Context Learning 
XL. Do*, **Y. Zhao***, H. Brown*, Y. Xie, X. Zhao, NF. Chen, K. Kawaguchi, M. Shieh, J. He, *ACL 2024 (Oral)*
- Felm: Benchmarking Factuality Evaluation of Large Language Models  
S. Chen, **Y. Zhao**, J. Zhang, IC. Chern, S. Gao, J. He, *accepted by NeurIPS 2023 D&B Track*
- Self-Evaluation Guided Beam Search for Reasoning  
Y. Xie, K. Kawaguchi, **Y. Zhao**, X. Zhao, MY. Kan, J. He, Q. Xie, *accepted by NeurIPS 2023*
- Joint Order Dispatch and Charging for Electric Self-Driving Taxi Systems 
G. Fan, H. Jin, **Y. Zhao**, Y. Song, X. Gan, J. Ding, L. Su, X. Wang, *accepted by INFOCOM 2022*
- Joint Order Dispatch and Repositioning for Urban Vehicle Sharing Systems via Robust Optimization 
Y. Zhao, G. Fan, H. Jin, W. Ma, B. He, X. Wang, *accepted by ICDCS 2021*
- Towards Fine-Grained Spatio-Temporal Coverage for Vehicular Urban Sensing Systems 
G. Fan, **Y. Zhao**, Z. Guo, H. Jin, X. Gan, X. Wang, *accepted by INFOCOM 2021*

PREPRINTS

- Identifying and Tuning Safety Neurons in Large Language Models
Y. Zhao, W. Zhang, Y. Xie, A. Goyal, K. Kawaguchi, M. Shieh, *submitted to ICLR 2025*
- From General to Expert: Custom Pruning LLMs Across Language, Domain, and Task
Y. Zhao, W. Zhang, G. Chen, K. Kawaguchi, L. Bing, *submitted to ICLR 2025*
- AdaMergeX: Cross-Lingual Transfer with Large Language Models via Adaptive Adapter Merging  
Y. Zhao, W. Zhang, H. Wang, K. Kawaguchi, L. Bing, *submitted to NAACL 2024*
- Investigating Pattern Neurons in Urban Time Series Forecasting
C. Wang, **Y. Zhao**, S. Cai, G. Tan, *submitted to ICLR 2025*

RESEARCH EXPERIENCE

Language Technology Lab in Alibaba DAMO Academy

Research Intern mentored by Dr. Wenxuan Zhang and Dr. Lidong Bing

Singapore

Aug. 2023 – Now

Topic: Training Base Models for SeaLLM 3: An Open Foundation and Chat Multilingual LLMs for Southeast Asian Languages

- We conduct efficient language enhancement by training language-specific neurons only based on a foundation model, significantly reducing the overall training cost. Moreover, such targeted training also ensures that the performance of high-resource languages can remain unaffected during the enhancement.
- Our model excels in tasks such as world knowledge, mathematical reasoning, translation, and instruction following, achieving state-of-the-art performance among similarly sized models.

Topic: From General to Expert: Custom Pruning LLMs Across Language, Domain, and Task

- We design a custom pruning method to prune a large general model into a smaller expert model for specific scenarios. It positions an expert model along the “language”, “domain” and “task” dimensions. By identifying and pruning irrelevant neurons, it creates expert models without any post-training.
- Our experiments demonstrate that custom pruning consistently outperforms other methods, achieving minimal loss in both expert and general capabilities across various models from different model families and sizes.

Topic: Understand How do Large Language Models Handle Multilingualism

- We introduce a novel method for detecting language-specific neurons within LLMs. Furthermore, we propose a new end-to-end framework for understanding the multilingual processing mechanism of LLMs.
- We find that language-specific neurons make up only 0.1%, meaning you can easily deactivate or enhance them.
- Moreover, LLMs predominantly “think” in English but draw upon multilingual world knowledge, with their self-attention and feed-forward layers respectively.

Topic: Cross-Lingual Transfer with Large Language Models via Adaptive Adapter Merging

- We propose a new cross-lingual transfer method called AdaMergeX that utilizes adaptive adapter merging.
- By introducing a reference task, we can determine that the divergence of adapters fine-tuned on the reference task in both languages follows the same distribution as the divergence of adapters fine-tuned on the target task in both languages. Hence, we can obtain target adapters by combining the other three adapters.
- Furthermore, we propose a structure-adaptive adapter merging method.

School of Computing in National University of Singapore

Ph.D. candidate supervised by Prof. Kenji Kawaguchi and Prof. Michael Shieh

Singapore

Nov. 2022 – Now

Topic: Identifying and Tuning Safety Neurons in Large Language Models

- Develop a neuron detection method to identify safety neurons—those consistently crucial for handling and defending against harmful queries.
- Introduce a safety neuron tuning method, that exclusively tune safety neurons without compromising models’ general capabilities. It significantly enhances the safety of instruction-tuned models can be applied to base models on establishing LLMs’ safety mechanisms.
- In addition, we improve the LLMs’ safety robustness during downstream tasks fine-tuning by separating the safety neurons from models’ foundation neurons.

Topic: Accelerating Greedy Coordinate Gradient via Probe Sampling

- Propose Probe Sampling to accelerate GCG, a universal and transferrable method to attack aligned LLMs.
- The core of the algorithm is a mechanism that dynamically determines how similar a smaller draft model’s predictions are to the target model’s predictions for prompt candidates.
- When the target model is similar to the draft model, we rely on the draft model to filter out a large number of potential prompt candidates to reduce the computation time.

Topic: Prompt Optimization via Adversarial In-Context Learning

- Revisit adversarial training in the context of large language models – given an NLP task, we propose to optimize the discrete prompts, a critical component for LLMs’ success, with an adversarial objective.
- Inspired by adversarial learning, adv-ICL is implemented as a two-player game between a generator and discriminator, with LLMs acting as both.

SELECTED HONORS

Zhiyuan Excellence Scholarship

2021, 2020, 2019, 2018