Zouying Cao

Telephone: 18362155818
Email: zouyingcao@sjtu.edu.cn
Github: github.com/zouyingcao
Website: zouyingcao.github.io
Google Scholar: zouyingcao

"I am an AI researcher. My wish is: efforts can be in proportion to gains!"

Education

`Sep. 2023-Present` **M.S. in Computer Science and Technology**, Shanghai Jiao Tong University.
Research Interests: Large Language Models (LLMs), Natural Language Processing (NLP)
GPA: 3.9/4.0, Advisor: Prof. Hai Zhao

`Sep. 2019-Jun. 2023` **B.S. in Computer Science and Technology**, Southeast University.
Outstanding Graduates of Southeast University, National Scholarship
GPA: 4.184/4.8, Ranking: 5/95, Advisor: Prof. Shuai Wang

Internships

`Jul. 2024-Present` **Alibaba**, Research Intern, Topic: LLM agents, primary focus on agent planning.

Publications

(* denotes equal contribution)

2025 **PGPO: Enhancing Agent Reasoning via Pseudocode-style Planning Guided Preference Optimization**.
Zouying Cao; Runze Wang; Yifei Yang; Xinbei Ma; Xiaoyong Zhu; Bo Zheng; Hai Zhao.
UnderReview | Code: zouyingcao/PGPO
Contribution: We investigate the effectiveness of pseudocode-style plans in agent reasoning, which are more concise and structured than NL plans. Based on two designed planning-oriented rewards, we further introduce PGPO, a preference optimization method that empowers LLM agents with enhanced reasoning capabilities.

LESA: Learnable LLM Layer Scaling-Up.
Yifei Yang; Zouying Cao; Xinbei Ma; Yao Yao; Libo Qin; Zhi Chen; Hai Zhao.
UnderReview | Code: yangyifei729/LESA | Paper: arxiv.2502.13794
Contribution: By applying SVD to concatenated parameters from each layer, we observe latent patterns such as continuity between layers, suggesting that inter-layer parameters can be learned. Therefore, LESA uses a neural network to predict the parameters inserted between adjacent layers for better LLM depth scaling-up.

2024 **KVSharer: Efficient Inference via Layer-Wise Dissimilar KV Cache Sharing**.
Yifei Yang; Zouying Cao; Qiguang Chen; Libo Qin; Dongjie Yang; Hai Zhao; Zhi Chen.
UnderReview | Code: yangyifei729/KVSharer | Paper: arxiv.2410.18517
Contribution: KVSharer is one layer-wise KV cache compression method based on the counterintuitive phenomenon where sharing dissimilar KV caches does not significantly degrade model performance.

SCANS: Mitigating the Exaggerated Safety for LLMs via Safety-Conscious Activation Steering.
Zouying Cao; Yifei Yang; Hai Zhao.
AAAI-2025-Oral (CCF-A) | Code: zouyingcao/SCANS | Paper: arxiv.2408.11491
Contribution: We investigate the safety defense mechanism by analyzing how the hidden states change when exposed to harmful queries and discover the extracted refusal steering vectors from middle layers promote refusal tokens (e.g., cannot), thereby steering the corresponding representation can reduce the false refusal rate.

Head-wise Shareable Attention for Large Language Models.
Zouying Cao; Yifei Yang; Hai Zhao.
EMNLP-2024-Findings (CCF-B) | Code: zouyingcao/DirectShare | Paper: arxiv.2402.11819
Contribution: We explore the feasibility of head-wise weight sharing across the attention heads in LLMs inspired by attention map (i.e., attention scores) reuse. Consequently, we propose two methods for head-wise weight sharing called DirectShare and PostShare, which are complementary in terms of time and performance.

LaCo: Large Language Model Pruning via Layer Collapse.
Yifei Yang; Zouying Cao; Hai Zhao.
EMNLP-2024-Findings (CCF-B) | Code: yangyifei729/LaCo | Paper: arxiv.2402.11187
Contribution: We propose a concise layer-wise structured pruner called Layer Collapse (LaCo), in which rear model layers collapse into a prior layer, enabling a rapid model size reduction while preserving the model structure.

2023 **AutoHall: Automated Hallucination Dataset Generation for Large Language Models**.
Zouying Cao; Yifei Yang; Hai Zhao.
UnderReview | Code: zouyingcao/AutoHall | Paper: arxiv.2310.00259
Contribution: We propose an approach called AutoHall for fast and automatically constructing model-specific hallucination datasets based on existing fact-checking datasets, eliminating the need for manual annotation.

Achievements

2024-11	Huatai Securities Technology Scholarship.
2024-11	First Prize Graduate Academic Scholarship in SJTU.
2023-06	Outstanding Graduates of Southeast University.
2023-06	**Outstanding Undergraduate Thesis Award of Southeast University.**
2022-10	Huawei Scholarship.
2021-12	**National Scholarship**.
2021-05	Second Price, National English Competition for College Students(NECCS)
2020-2022	Three-good Student for three consecutive years
2020-12	Scholarship on Social Works in SEU.
2020-12	**President Scholarship in SEU.**
2020-05	Excellent League Cadres, Award for the Models of the Chinese Youth in SEU.
2020-05	Third Prize, the 17th Southeast University College Student Programming Competition.

University Projects

Sep.2022–Dec.2022	**Embedded Computer System Minisys-1A SoC** | Code: zouyingcao/minisys-1A Role: Leader, Hardware part, Verilog
Apr.2022–May.2022	**A B2C Hotel Booking Platform Called Hippo Hotel** | Code: zouyingcao/hotel_bankend *A B2C online website to support both hotel reservation and hotel management.* Role: Back-end developer, SpringBoot & MySQL
Nov.2020–Nov.2021	**A BiLSTM-based Travel Time Estimation Model** | Code: zouyingcao/ETA_Project *A data-driven deep learning model based on BiLSTM to estimate time of arrival.*

Patents

2023-09-26	**A Multi-sided Fairness-Aware Order Dispatch Method for Instant Delivery** | CN202310387578.6
2023-05-09	**A Gesture Recognition and Tracking Method and System for Mobile Devices** | CN202211488944.9

Miscellaneous

Reviewers	ARR (ACL Rolling Review)	
Skills	CET6-568/CET4-643, familar with Python, C++, data structure and algorithm	
Social Works	Sep 2023-Present	Party branch secretary, in SJTU
	Aug 2021-Jun 2023	2021 undergraduate class instructor, in SEU
	Sep 2020-Jun 2023	Red Cross Society of China Nanjing Branch, Member
	Sep 2019-Sep 2020	League branch secretary & secretary of student union, in SEU
Volunteering	volunteer for freshmen, first-aid personnel, volunteer during the COVID-19 pandemic	
Interests	painting, music, and delicious food	