Zefan Cai

zefncai@gmail.com · Google Scholar ·

Education

University of Wisconsin - Madison - School of Computer 2024 – Present

Ph.D. majoring in Computer Science
Supervisor: Prof. **Junjie Hu** from School of Computer, Data & Information Sciences

Peking University - School of Software & Microelectronics 2022 – 2024

M.Sc. majoring in Computer Science Technology; GPA: 3.71/4.0 (top 5%)
Supervisor: Prof. **Baobao Chang** from Institute of Computational Linguistics (ICL) from Peking University

Beijing Jiaotong University - College of Computer Science and Technology 2018 – 2022

B.Eng. majoring in Computer Science; GPA: 3.61/4.0 (Rank 2/31)

Publications, Submissions and Preprints

My previous research mainly include **KV Cache Comppression** [1] [2] [12], **Efficiency** [1] [2] [8] [12], **Instruction Tuning** [3] [4] [5], **Reasoning** [14] [17], **Evaluation** [10], **Preference Learning** [21], **Agent** [15] [18], **Vision-Language Understanding** [4] [8] [16] [18], **Vision-Language Generation** [13], and **Traditional NLP** [6] [7] [9] [20].

1. KVCache-Factory: Unified KV Cache Compression Methods for Auto-Regressive Models [Code]
   Zefan Cai **Open-Source Project**
2. PyramidKV: Dynamic KV Cache Compression based on Pyramidal Information Funneling [PDF] [Code]
   Zefan Cai, Yichi Zhang, ..., Baobao Chang, Junjie Hu, Wen Xiao **Preprint**
3. Improving Event Definition Following For Zero-Shot Event Detection [PDF] [Code]
   Zefan Cai*, Po-Nien Kung*, ..., Wei Wang, Nanyun Peng **ACL 2024 Main Conference Long Paper**
4. MMICL: Empowering VLM With Multi-Modal In-Context Learning [PDF] [Code] [Data] [Model]
   Haozhe Zhao*, **Zefan Cai***, ..., Zixuan Liu Sheng Wang, Wenjuan Han, Baobao Chang **ICLR 2024, Poster**
5. Compositional Task Representations for Large Language Models [PDF] [Code]
   Nan Shao*, **Zefan Cai***, Chonghua Liao, Yanan Zheng, and Zhilin Yang **ICLR 2023 Poster**
6. DialogVCS: Robust Natural Language Understanding in Dialogue System Upgrade [PDF]
   Zefan Cai*, Xin Zheng*, ..., Baobao Chang, Yunbo Cao **NAACL 2024 Main Conference Long Paper**
7. Mitigating Language Performance Disparity in multi-lingual Pre-trained Language Models via Teacher Language Selection and Cross-lingual Distillation [PDF] [Code]
   Haozhe Zhao*, **Zefan Cai***, Shuzheng Si, ..., Baobao Chang **NAACL 2024 Main Conference Long Paper**
8. VeCAF: VLM-empowered Collaborative Active Finetuning with Training Objective Awareness [PDF]
   Rongyu Zhang*, **Zefan Cai***, ..., Kurt Keutzer, Baobao Chang, ..., Shanghang Zhang **ACMMM 2024**
9. SANTA: Separate Strategies for Inaccurate and Incomplete Annotation Noise in DS-NER [PDF] [Code]
   Shuzheng Si*, **Zefan Cai***, ..., Jiaxing Lin, Baobao Chang **ACL 2023 Findings Long Paper**
10. Large Language Models are not Fair Evaluators [PDF] [Code]
    Peiyi Wang, Lei Li, Liang Chen, **Zefan Cai**, ..., Zhifang Sui **ACL 2024 Main Conference Long Paper**
11. PCA-Bench: Evaluating Multimodal Large Language Models in Perception-Cognition-Action Chain [PDF] [Code]
    Liang Chen, ..., **Zefan Cai**, ..., Baobao Chang **ACL 2024 Findings Long Paper**
12. Not All Heads Matter: A Head-Level KV Cache Compression Method with Integrated Retrieval and Reasoning [PDF]
    Yu Fu, **Zefan Cai**, ..., Wen Xiao **Preprint**
13. A Spark of Vision-Language Intelligence: 2-Dimensional Autoregressive Transformer for Efficient Fine-grained Image Generation [PDF]
    Liang Chen, Sinan Tan, **Zefan Cai**, ..., Baobao Chang **Preprint**
14. Human-In-The-Loop through Chain-of-Thought [PDF]
    Zefan Cai, Baobao Chang, Wenjuan Han **Preprint**
15. ML-Bench: LLMs Leverage Open-source Libraries for Machine Learning Tasks [PDF] [Code] [Page]
    Xiangru Tang*, Yuliang Liu*, **Zefan Cai***, ..., Baobao Chang, ..., Arman Cohan, Mark Gerstein **Preprint**
16. DiffCap: Exploring Continuous Diffusion on Image Captioning [PDF] [Code]
    Yufeng He*, **Zefan Cai***, Xu Gan, Baobao Chang **Preprint**

17. LLM Critics Help Catch Bugs in Mathematics: Towards a Better Mathematical Verifier with Natural Language Feedback [PDF]
    Bofei Gao, **Zefan Cai**, ... Baobao Chang **Preprint**
18. COMMA: A Communicative Multimodal Multi-Agent Benchmark [PDF]
    Timothy Ossowski, ..., **Zefan Cai**, ..., Junjie Hu **Preprint**
19. Omni-math: A universal olympiad level mathematic benchmark for large language models [PDF]
    Bofei Gao, Feifan Song, Zhe Yang, **Zefan Cai**, ... **Preprint**
20. CENSOR: Distantly-Supervised Named Entity Recognition with Uncertainty-aware Teacher Learning and Student-student Collaborative Learning [PDF]
    Helan Hu, ..., **Zefan Cai**, Baobao Chang **ACL 2024 Findings Long Paper**
21. Towards a Unified View of Preference Learning for Large Language Models: A Survey [PDF] [Code] [Page]
    Bofei Gao, Feifan Song, Yibo Miao, **Zefan Cai**, ... Baobao Chang **Preprint**

Research Projects

University of Wisconsin - Madison - Supervisor: Junjie Hu Sept. 2024 – Now

KVCache-Factory: Unified KV Cache Compression Methods for Auto-Regressive Models [1]

- Create a efficient KV cache optimization library for Transformer-based foundation models.
- Include the unified implementation of KV cache compression methods at different stages and different categories.
- Include acceleration methods at Prefilling stage, KV cache management stage and Decoding Stage.
- For KV cache management, include KV cache compression methods such as Merge, Quantization and Eviction.

PyramidKV: Dynamic KV Cache Compression based on Pyramidal Information Funneling [2]

- Observe **Pyramidal Information Funneling** that LLMs aggregate information where attention is scattering widely in lower layers, and ultimately focusing on critical tokens in higher layers.
- Develope **PyramidKV**, which dynamically adjusts the KV cache size across different layers.
- Experiments indicate that PyramidKV shows superior performance, especially under resource-intensive circumstances and effectively preserves long context capabilities.
- Implementations show that PyramidKV can be integrated with pre-filling stage methods (e.g., MInference) or other KV cache management methods (e.g., KV cache quantization) and diverse frameworks (e.g., Huggingface and vLLM).

University of California - Los Angeles - Supervisor: Nanyun Peng June. 2023 – Sept. 2023

Improving Event Definition Following For Zero-Shot Event Detection [3]

- Explore whether LLM-based Event Detection models can **generalize to unseen events** by given event definitions.
- Evalaute **scaling law on EE** by training LLM in scales (i.e. # event type , # sample and # event definition).
- Fine-tune LLM with LLM-generated event types & event definitions & samples and evaluate in unseen event types.

Honors and Awards

Merit Student (Top 10%), Peking University Sept. 2023
Merit Student (Top 10%), Beijing Jiaotong University Sept. 2020

Academic Service

- Reviewer: EMNLP, ICML, ICLR, Neurips, ARR