Zhaorui Yang

zhaorui.yang@zju.edu.cn https://rickyang1114.github.io rickyang1114

Education

Zhejiang University — Hangzhou, China
Ph.D. student in Software Engineering — *Sept. 2023 – Present*

Xi'an Jiaotong University — Xi'an, China
B.E. in Software Engineering — *Sept. 2019 – June 2023*

Publications

- **Zhaorui Yang***, Bo Pan*, Han Wang*, Yiyao Wang, Xingyu Liu, Luoxuan Weng, Yingchaojie Feng, Haozhe Feng, Minfeng Zhu, Bo Zhang, Wei Chen. Multimodal DeepResearcher: Generating Text-Chart Interleaved Reports From Scratch with Agentic Framework. **AAAI 2026 (Oral)**.

  Existing deep research frameworks primarily focus on generating text-only content, leaving the automated generation of interleaved texts and visualizations underexplored. In this work, we propose Formal Description of Visualization (FDV), a structured textual representation of charts that enables LLMs to learn from and generate diverse, high-quality visualizations. Building on this representation, we introduce Multimodal DeepResearcher, an agentic framework that automatically generates comprehensive multimodal reports from scratch with interleaved texts and visualizations.

- **Zhaorui Yang**, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, Qian Liu. Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning. **ACL 2024**.

  Fine-tuning LLMs for specific tasks often encounters challenges in balancing performance and preserving general instruction-following abilities. In this work, we posit that the distribution gap between task datasets and the LLMs serves as the primary underlying cause. To address the problem, we introduce Self-Distillation Fine-Tuning (SDFT), a novel approach that bridges the distribution gap by guiding fine-tuning with a distilled dataset generated by the model itself to match its original distribution.

- Haozhe Feng*, **Zhaorui Yang***, Hesun Chen*, Tianyu Pang, Chao Du, Minfeng Zhu, Wei Chen, Shuicheng Yan. CoSDA: Continual Source-Free Domain Adaptation. arXiv preprint arXiv:2304.06627, 2023.

  In this work, we investigate the mechanism of catastrophic forgetting of previous Source-Free Domain Adaptation (SFDA) approaches. We observe that there is a trade-off between adaptation gain and forgetting loss. Motivated by the findings, we propose CoSDA, which outperforms SOTA approaches in continuous adaptation.

Internships

Tencent | *Data Computing Platform Department, TEG* (In progress) — July 2025 – Present

- Developing an LLM-powered Data Agent to automate "Text-to-Insight" analysis across enterprise data warehouses containing 1,000+ tables per user, enabling natural language queries for complex data exploration.
- Designed and implemented profiling algorithms to handle massive-scale tables with heterogeneous data types, including semi-structured formats (JSON, ARRAY) and nested structures.
- Built end-to-end analytical pipeline encompassing schema linking, automated SQL generation, and Python-based data analysis, streamlining the workflow from user intent to actionable insights.

Awards

China National Scholarship (Undergraduate). — *Dec. 2022*

China National Scholarship (Undergraduate). — *Dec. 2021*