Zhe Wang

zhewang2001.github.io zhe36@illinois.edu github.com/zhewang2001

Education

University of Illinois at Urbana-Champaign **Aug 2024 – May 2026**
Siebel School of Computing and Data Science *Illinois, United States*
- Master of Science in Computer Science
- **Advisor:** Lingming Zhang

Tsinghua University **Sep 2020 – Jun 2024**
Weiyang College *Beijing, China*
- Bachelor of Science in Mathematics and Physics
- **Advisor:** Zhiyuan Liu

Research Interests

- **Transfer Learning:** to have a more fundamental understanding of the behavior of large language models (LLMs), enabling their effective adaptation to knowledge-intensive and everyday scenarios.
- **AI for Code:** to build and evaluate AI programmers to propel developer efficiency and improve software quality.

Publications

- Yuxiang Wei, **Zhe Wang**, Jiawei Liu, Yifeng Ding, Lingming Zhang. *"Magicoder: Empowering code generation with oss-instruct"* ICML 2024.(https://arxiv.org/abs/2312.02120)

Research Experiences

Magicoder: Source Code Is All You Need **Oct 2023 – Dec 2023**
Research Intern at Illinois at Urbana-Champaign, Supervised by Prof. Lingming Zhang *Beijing, China*
- Aimed to mitigate the inherent bias of the synthetic data generated by LLMs by empowering them with a wealth of open-source references for the production of more diverse, realistic, and controllable data.
- Proposed a new approach, OSS-INSTRUCT, to enlighten LLMs with open-source code snippets to generate high-quality instruction data for code. Magicoder, our model trained on 75K synthetic instruction data using OSS-INSTRUCT, can outperform all evaluated LLMs with less than or equal to 16B parameters despite having less than 7B parameters.
- Achieved more than 2k stars on Github and was on Github trending.

ProgEval: An Evaluation Benchmark For Code Generation **Jul 2023 – Aug 2023**
Research Intern at Illinois at Urbana-Champaign, Supervised by Prof. Lingming Zhang *Beijing, China*
- Developed a structured ProgEval problem format consisting of a detailed description, entry and key function signatures, with associated tests, and provided helper function implementations.
- Creating a manual CodeGen benchmark, comprising 100 Python programming problems across diverse topics and dependencies.
- Designed and implemented an evaluation pipeline, a generic model inference interface compatible with OpenAI, Google, and Huggingface, and an adaptive prompt builder. Additionally, established a comprehensive metrics and evaluation loop for thorough assessment.
- Evaluating auxiliary functions' effects on LLMs' problem-solving and tool-using capabilities.

Catastrophic Forgetting in Continual Instruction Fine-tuning **May 2023 – Present**
Research Intern at Wisconsin-Madison, Supervised by Prof. Junjie Hu *Beijing, China*

- Provided a comprehensive study in what and how instruction-finetuning LLMs(IF-LLMs) forget from the perspective of data distribution and how to mitigate catastrophic forgetting in IF-LLMs.
- Quantified data distribution of different training data and conducted detailed experiments on how IF-LLMs suffer from catastrophic forgetting in different data distributions.
- Adapted traditional methods for mitigating catastrophic forgetting to suit IF-LLMs, including regularization-based methods and replay-based methods.
- Prepared to submit a paper to ACL 2024 as the first author.

Diffusion-Optimizer Speeds Up Delta Tuning **Sept 2022 – Mar 2023**
Research Intern at Tsinghua University, Supervised by Prof. Zhiyuan Liu *Beijing, China*

- Aimed to achieve better performance beyond traditional optimizers by using G.pt (Generative pre-training from checkpoints) model based on diffusion model principle for delta tuning of training data.
- Conducted G.pt training with T5-prompt-tuning checkpoints pretraining data on SST-2 task.
- Enhanced optimization ability of diffusion-optimizer through data augmentation, modifying the model structure, and other methods.

A Q&A System Based on Human Search Engine Behavior Modeling **Jul 2022 – Aug 2022**
Research Intern at Tsinghua University, Supervised by Prof. Zhiyuan Liu *Beijing, China*

- Aimed to train AI models to generate questions,search behaviors and desired answers intelligently with data on human behavior when using search engines for long-text knowledge-based question answering.
- Built a web-based search engine platform and validated the effectiveness of the pipeline using CPM1(Chinese Pre- training Model) model.
- Participated in collecting high-quality search behavior data for 6000 questions, labeled and processed data.
- Helped to set up the framework and some of the pipeline code for the CPM1 normalsize model.

Technical Skills

Languages: Mandarin (Native), English (TOEFL 105: **Speaking 25**), Japanese Beginner
Programming Languages: Python, Java, C, HTML/CSS, JavaScript, SQL
Developer Tools: VS Code, Eclipse, Google Cloud Platform, Jupyter Notebook
Technologies/Frameworks: Linux, GitHub, Huggingface, Tensorflow, Pytorch