Shuyang Jiang

(+86) 17721315376
https://pixas.github.io/
jiangshuyang0@gmail.com

*PhD students at Fudan University, interested in machine learning, natural language processing and AI in medicine. I have been involved in research for 5 years since the junior year in my undergraduate period, with **157** citations in total*

Research EXPERIENCE

Finetuning with Reserved Majority for Noise Reduction **Shanghai**
First Author; project lead; [ICLR 2025 Spotlight] *July 2024–Sept 2024*

- *Discover common redundancies among parameter-efficient fine-tuning (PEFT) scenarios in modern large language models*
- *Propose NoRM, a lightweight post-processing technique to reduce such redundancies through sub-space similarity with base weights, achieving ~5 points improvements with LoRA under no extra inference latencies*

Taia: Large language models are out-of-distribution data learners **Shanghai**
First Author; project lead; [NeurIPS 2024 Poster] *Feb 2024–May 2024*

- *Unveil that in Low-rank Adaptor (LoRA), it is essential to fine-tune all parameters but remain only the attention part (TAIA), to obtain strong out-of-distribution generalization*
- *We validate this finding across four backbone models, under various training data and testbeds. TAIA shows superior performance gains over LoRA baselines.*

MedS^3: Towards Medical Small Language Models with Self-Evolved Slow Thinking **Shanghai**
First Author; project lead; [Arxiv Preprint] *Sept 2024–Jan 2025*

- *Use Monte-Carlo Tree Search to conduct exploration on ~8000 seed clinical instances and train a process-level supervision reward model as well as a medical reasoning policy model for test-time scaling*
- *The policy and process reward model system outperform the strongest open-source baseline by 8 points in a testbed covering 11 benchmarks overall, showcasing the high efficiency and efficacy of proposed medical reasoning system.*

Attentive Multi-Layer Perceptron for Non-autoregressive Generation **Shanghai**
First Author; [ECML-PKDD 2023] *Jan 2023–May 2023*

- *Aimed at improving efficiency of non-autoregressive transformers (NAT), we propose Attentive Multi-layer Perceptron (AMLP) layers to replace vanilla attention. It decouples the QK production in attention and computes KV first to reduce quadratic computation*
- *AMLP achieves linear time complexity with 1 point improvement on WMT 14 de-en.*

EDUCATION

Fudan University **Shanghai**
Ph.D in computer science (concentration: computer science), *2023.09-present*

Shanghai Jiao Tong University **Shanghai**
Bachelor of computer science (concentration: computer science), *2019.09-2023.06*
Honors: Outstanding graduates (GPA: 3.96/4.30)