LIJUN WU

(+86) 15901525751 apeterswu@gmail.com https://apeterswu.github.io

SHORT BIO

Lijun Wu is currently a Research Scientist in Shanghai AI Laboratory. Previously, he was a Research Scientist in ByteDance, a Senior Researcher in MSR AI4Science/Microsoft Research Asia. He got the Ph.D. degree from Sun Yat-sen University (SYSU) in 2020 and was a member of joint Ph.D. program between SYSU and MSRA, advised by Dr. Tie-Yan Liu and Prof. Jianhuang Lai. He received MSRA Ph.D. Fellowship in 2018. His research focuses on LLM, Data-centric AI, NLP, and AI4Science.

EDUCATION

Sun Yat-Sen University *Sep. 2015 to Jun. 2020*
Ph.D. in Computer Science and Technology◇ Joint Ph.D. Program with MSRA
School of Data and Computer Science
Ph.D. Supervisor: Tie-Yan Liu and Jianhuang Lai

EXPERIENCES

- **ByteDance/Seed LLM** *May. 2024 to Aug. 2024*
  Research Scientist ◇ *Large Language Modeling*
- **Microsoft Research Asia/AI4Science** *Jun. 2020 to May. 2024*
  Senior Researcher ◇ *Deep Learning*
- **Microsoft Research Asia** *Jan. 2014 to Jun. 2020*
  Research Intern ◇ *Machine Learning Group* **Mentor: Tao Qin, Tie-Yan Liu**

SELECTED PUBLICATIONS

Full publications in Google Scholar

- **LLMs (post-training):**
  - **InternVL3, InternVL3.5 (foundation models)**
  - **ScaleDiff, Caco, GRA (data synthesis/reasoning)**
- **AI/NMT:**
  - **R-Drop, UniDrop (consistency training)**
  - **RL4NMT, BERT-NMT, Mono-NMT (neural machine translation)**
- **AI4Science:**
  - **BioT5, BioT5+, 3D-MolT5 (pre-trained LLM for text and bio-chemistry)**
  - **FABind, FABind+, FABFlex (Fast and accurate Protein-Ligand Binding)**
  - **NatureLM (scientific foundation model), μFormer (protein engineering), TamGen (drug design)**

KEY PROJECTS

OpenDataArena: Benchmark Data Value in LLM *Mar. 2025 to now*

- I lead the development of OpenDataArena (ODA), an open and transparent platform for systematically evaluating the value of post-training datasets for large language models (LLMs). The project addresses a critical gap in the field—while model architectures advance rapidly, dataset quality remains underexplored despite being a key driver of performance. I designed and implemented the multi-dimensional scoring framework, established the training–evaluation pipeline for models such as LLaMA and Qwen, and led the analysis of over 100 datasets across domains including general instruction, math, code, science. ODA has processed more than 40 million data points and generated reproducible benchmarks that reveal meaningful relationships between data quality and downstream model performance.

Multilingual Low-resource Large Language Model *Oct. 2024 to Jan. 2025*

- I led a team of 12 to develop a specialized multilingual LLM focused on low-resource Hungarian, Chinese, and English. By leveraging advanced techniques like data synthesis and continued training on Qwen2.5, our model achieved state-of-the-art results, surpassing GPT-4o's performance in understanding Hungarian culture, politics, and knowledge. The project concluded with the successful deployment of the model at a key partner organization.

Drug Discovery and Scientific Foundation Model *Oct. 2021 to May. 2024*

- We aim to build a groundbreaking project aimed at building a versatile scientific model, NatureLM, for tackling challenges in drug discovery, materials science, and biology. My primary focus was on drug-target binding prediction tasks, including interactions, affinities, and molecular docking. Collaborating with an interdisciplinary team, I helped develop and implement machine learning algorithms for accurate predictions, contributing to the project's success and expanding my expertise in computational drug discovery. *Several techniques have been transferred into the Microsoft Azure platform to empower customers.*

WMT 2019 Machine Translation Competition *Feb. 2019 to Mar. 2019*

- Our team participate in the WMT 2019 machine translation competition in 11 translation directions, and we obtain 1st place in 8 translations and 2nd place in other 3 translations. Specifically, I participate in 5 translations: English-German, German-English, German-French, French-German and Russian-English, and we achieve *1st place in all the 5 translation directions, with more than 1.0 BLEU better than 2nd in the first four translations.* I am the main member in this project, I contribute data filtering, data usage in a scientific way, transdctive distillation technology, soft contextual argumentation technology, multi-agent dual learning technology and experiment running in the project.

Human Parity on Neural Machine Translation *Oct. 2017 to Mar. 2018*

- *Neural Machine Translation* (NMT) is proven to significantly outperform traditional translation techniques. To further improve the accuracies of NMT, we explore different structures (deep models, deliberation networks, efficient group network), attention mechanism design (word attention), training loss (sequence level loss), that can boost the performances of NMT from various aspects. *Specifically, our system firstly matches the human translation accuracy in Chinese-to-English machine translation in Mar. 2018.* I am one of the main contributor of this project, including algorithm design, system implementation and working on experiments.

HONORS & AWARDS

1. 2^{nd} in Internal Reasoning Track of CURE-Bench@NeurIPS2025 *2025*
2. 1^{st} of Text2Mol and 2^{nd} of Mol2Text in Language+Molecules@ACL2024 *2024*
3. Runner up of OGB-LSC@KDD cup *2021*

4. **Outstanding Graduate Awards of SYSU** *2020*
5. **Outstanding Reviewer of EMNLP** *2019*
6. **1^{st} of WMT 2019 machine translation competition in 5 translations** *2019*
7. **Microsoft Research Asia Ph.D. Fellowship** *2018*
8. **Graduate Student National Scholarship** *2018*
9. **Outstanding Graduate Awards** *2015*
10. **1^{st} of IBM/IEEE Smarter Planet Challenge** *2013*
11. **Undergraduate Student National Scholarship** *2012, 2013*

ACTIVITIES

Academic Serving:

- **AC: ICML-26, ICLR-26, NeurIPS-25, ACL-21/Now, EMNLP-23/Now, NAACL-22/Now, EACL-24, ARR-21/Now**
- **SPC: AAAI-22/Now, IJCAI-21**
- **PC member: ICLR, ICML, NeurIPS, AAAI, IJCAI, CVPR, ACL, KDD, EMNLP, NAACL, COLING, EACL, AACL**
- **Journal reviewer: TPAMI, TASLP, Neurocomputing, KBS, CSL, TALLIP**