

LIJUN WU

☎ (+86) 15901525751 ✉ lijun_wu@outlook.com 🌐 apeterswu.github.io

SHORT BIO

Lijun Wu is a **Young Scientist** at **Shanghai AI Laboratory** and an Adjunct Ph.D. Supervisor at **Shanghai Jiao Tong University**, **Fudan University**, and **Zhongguancun Academy**. He received his Ph.D. from Sun Yat-sen University through the joint Ph.D. program with **Microsoft Research Asia (MSRA)**, advised by **Dr. Tie-Yan Liu** and **Prof. Jianhuang Lai**. He was a Research Scientist at **ByteDance Seed** and a Senior Researcher at **Microsoft Research/AI4Science**.

His research spans LLMs, MLLMs, Data-centric AI, and AI4Science. He has about **10,000 citations**. Open-source projects he **leads** have **400k+ HuggingFace downloads**; models from flagship team efforts he **contributed**—**MinerU 2.5/InternVL series/Intern-S1 series**—collectively reach **multi-million** HuggingFace downloads. His research is deployed in industry at scale. He received the prestigious **MSRA Ph.D. Fellowship** in 2018 and is **Track Chair of NeurIPS-2026** (Evaluations & Datasets).

RESEARCH IMPACT

- **Citations:** ~10,000 ([Google Scholar](#)); publications in *Nature Communications*, *Nature Machine Intelligence*, *NeurIPS*, *ICML*, *ICLR*, *ACL*, *KDD*, *TPAMI*.
- **Open-source models & datasets:** **BioT5 series 300k+ downloads**; **OpenDataArena / ODA-Math / ODA-Mixture / MMFineReason datasets 100k+ downloads**; among contributed team projects, **InternVL (10k stars)** and **MinerU (57k+ stars)**; multiple projects reached **HuggingFace Trending Top 1–2**.
- **Industrial deployment:** **R-Drop** integrated into **Microsoft Translator** (20+ language directions) and Meituan, Wechat, Baidu's online systems; drug-discovery techniques transferred to **Microsoft Azure** platform.
- **Competition:** **8 championships** at **WMT-2019**; **1st Text2Mol & 2nd Mol2Text** (**ACL-2024 Language+Molecules**); **2nd NeurIPS-2025 CURE-Bench Internal Reasoning** track.

OPEN-SOURCE HIGHLIGHTS

- **InternVL3 / InternVL3.5 / Intern-S1-Pro** — Participated in the InternVL series and Intern-S1 series at Shanghai AI Lab, one of the most competitive open-source (scientific) VLM families.
- **MinerU 2.5 (Report)** — Participated in this widely used open-source toolkit for PDF-to-structured-text extraction (**57k+ GitHub stars**); contributions focused on chart and figure parsing and related multimodal understanding.
- **OpenDataArena (ODA)** — The first open arena for benchmarking post-training data value for LLMs/MLLMs; released ODA related datasets, **HuggingFace Trending Top 1–2**.
- **BioT5 / BioT5+** — Open-source pre-trained LLMs for unified text–molecule–biology understanding; **300k+ downloads** on HuggingFace; won 1st (Text2Mol) and 2nd (Mol2Text) at the **ACL-2024 Language+Molecules** shared task.
- **MMFineReason** — A 4B VLM achieving 30B-level performance via fine-grained reasoning; associated 1.8M-sample dataset reached **HuggingFace Trending Top 2**.
- **R-Drop** — Simple yet powerful regularization technique; deployed in **Microsoft Translator** across 20+ language pairs and widely adopted by the community, Wechat, Meituan, Baidu, etc.

KEY PROJECTS

InternVL3 / InternVL3.5 / Intern-S1-Pro: Open-Source VLMs

2025 to now

- Participated in the development of the InternV/Intern-S1 series of open-source VLMs at Shanghai AI Laboratory. InternVL3, InternVL3.5, and Intern-S1-Pro achieve top-tier performance across a broad spectrum of vision-language benchmarks, demonstrating the power of open, community-driven model development. The project demonstrates that open-source models can match or surpass proprietary systems, advancing both the research community and downstream applications.

MinerU 2.5: Open-Source Document Parsing

2025 to now

- Participated in the development of **MinerU 2.5**, an open-source vision-language pipeline for converting complex PDFs into LLM-ready markdown and structured outputs. My work emphasized chart and figure parsing—improving recognition and layout understanding for plots, diagrams, and mixed graphical content alongside broader document elements.

OpenDataArena: Benchmarking Post-Training Data Value

Mar. 2025 to now

- Led the development of **OpenDataArena (ODA)**, the first open and transparent platform for systematically evaluating the value of post-training datasets. The project released the multi-dimensional scoring framework, established the training-evaluation pipeline for open-source models, and analyzed 100+ datasets across different domains covering 40M+ data points. ODA has become a community resource for data-centric LLM research.

Multilingual Low-Resource Large Language Model

Oct. 2024 to Jan. 2025

- Led a team to develop a specialized multilingual LLM for low-resource Hungarian, Chinese, and English. By leveraging data synthesis and continued pre-training on Qwen2.5, our model surpassed GPT-4o's performance on Hungarian cultural, political, and general knowledge benchmarks. The model was successfully deployed at a key partner organization.

Scientific Foundation Models

Oct. 2021 to May 2024

- Lead the development of **BioT5** series models and co-lead **NatureLM**, the scientific foundation model tackling drug discovery, biology, and materials science. The foundation models achieve SOTA results on multiple tens of scientific tasks. The work also produced multiple high-impact papers including **FABind**, **FABind+**, and **TamGen**. Selected techniques have been transferred to the **Microsoft Azure** platform.

WMT 2019 Machine Translation Competition

Feb. 2019 to Mar. 2019

- Led the core technical effort for 5 translation directions (En-De, De-En, De-Fr, Fr-De, Ru-En), achieving **1st place in all 5 directions**, with >1.0 BLEU improvement over 2nd place (Meta). Key contributions: data filtering, transductive distillation, soft contextual augmentation, and multi-agent dual learning. Overall the team achieved **1st in 8 and 2nd in 3** of 11 competing directions.

Human Parity on Chinese-to-English Machine Translation

Oct. 2017 to Mar. 2018

- Contributed to the landmark Microsoft project that **first achieved human-parity accuracy** in Chinese-to-English news translation. Designed and implemented improvements across deep model architectures, deliberation networks, attention mechanisms, and sequence-level training objectives.

SELECTED PUBLICATIONS

Full list: [Google Scholar](#)

- **LLMs / MLLMs — Foundation Models & Post-Training:**

- [InternVL3](#), [InternVL3.5](#), [Intern-S1-Pro](#) — open-source vision-language foundation models
- [MMFineReason](#), [ScaleDiff](#), [CaCo](#) — data synthesis and reasoning for post-training
- [MinerU 2.5](#) — open-source PDF/document parsing

- **Scientific Foundation Models/AI4S:**

- [BioT5](#), [BioT5+](#), [3D-MolT5](#) — scientific foundation model, 300k+ HuggingFace downloads
- [NatureLM](#) — scientific foundation model
- [\$\mu\$ Former](#) (*Nature Machine Intelligence*) — protein engineering
- [TamGen](#) (*Nature Communications*) — target-aware molecule generation for drug design

- **Neural Machine Translation:**

- [RL4NMT](#) — among the first studies of RL for sequence generation
- [BERT-NMT](#), [Mono-NMT](#) (backed WMT-2019 champion system)
- [NAT Survey](#) — comprehensive survey on non-autoregressive generation

EXPERIENCE

- **Shanghai AI Laboratory** *2024 – Present*
Young Scientist \diamond *Large Language Models, Multimodal LLMs, Data-centric AI*
Adjunct Ph.D. Supervisor: Shanghai Jiao Tong University, Fudan University, Zhongguancun Academy
- **ByteDance / Seed LLM** *2024 – 2024*
Research Scientist \diamond *Large Language Model post-training*
- **Microsoft Research AI4Science / MSRA** *2020 – 2024*
Senior Researcher \diamond *AI4Science, Deep Learning, NLP*
- **Microsoft Research Asia** *2014 – 2020*
Research Intern \diamond Machine Learning Group
Mentors: [Tao Qin](#), [Tie-Yan Liu](#)

SELECTED HONORS & AWARDS

1. 2nd place in Internal Reasoning Track, [CURE-Bench@NeurIPS-2025](#) *2025*
2. [Top Area Chair](#), [NeurIPS-2025](#) *2025*
3. 1st place in Text2Mol & 2nd place in Mol2Text, [Language+Molecules@ACL-2024](#) *2024*
4. Runner-up, [OGB-LSC@KDD Cup 2021](#) *2021*
5. 1st place in 8 directions, [WMT-2019 Machine Translation Competition](#) *2019*
6. [Microsoft Research Ph.D. Fellowship](#) *2018*
7. Stars of Tomorrow Internship Award of MSRA *2018*
8. 1st place of [Global IBM/IEEE Smarter Planet Challenge](#) *2013*

ACADEMIC SERVICE

- **PC:** Evaluations & Datasets Track Chair of [NeurIPS-2026](#)
- **AC:** [ICML-26](#), [ICLR-26](#), [KDD-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](#)
- **Journal Reviewer:** [TPAMI](#), [TASLP](#), [Neurocomputing](#), [KBS](#), [CSL](#), [TALLIP](#)