

Wei Li, PhD

AI researcher and engineer. Compression, adaptation and evaluation of large language models.

London, United Kingdom | wei@wgtech.co.uk | wgtech.co.uk
[Google Scholar](#) | [OpenReview](#) | [ORCID 0000-0002-4776-5911](https://orcid.org/0000-0002-4776-5911)

Profile

I work on making large language models smaller and cheaper to adapt. My PhD at the University of Birmingham covered pruning, low-rank compression and knowledge distillation, with first-author papers at NeurIPS 2024, ICML 2025 and Findings of ACL 2025.

In 2026 I founded WGTECH LTD, a London company that builds AI workflow and office automation, assistants, and web and iOS applications. I lead scoping and the AI side of each project, alongside colleagues who cover backend, web and iOS development.

Experience

WGTECH LTD Founder and Director, London Jun 2026 – present

AI engineering company registered in England and Wales (no. 17304976). We build workflow and office automation, assistants that answer from a client's own documents, and web and iOS applications, and we tune models when an off-the-shelf one is too slow or too costly.

- Lead project scoping, workflow and model design, evaluation and technical direction.
- Work with a second AI engineer and with backend (Spring), web (Vue) and iOS (Swift) developers.

ZoyMed Co-Founder and Technical Lead, Shenzhen Apr 2025 – present

Medical AI company working on surgical video analysis for education. The system identifies key moments in procedure recordings, which surgeons use to prepare teaching material, conference talks and patient consultations.

Jilin University Visiting Associate Professor, Changchun Oct 2024 – present

Research collaboration with Prof. Xuan Song at the Key Laboratory of International Technology Cooperation for Super-Intelligent Cities. The appointment runs to October 2029.

University of Birmingham Teaching Assistant, Birmingham Sep 2020 – May 2023

Supported teaching on Programming in C/C++ and Artificial Intelligence 1.

Education

PhD, Computer Science University of Birmingham 2020 – 2026

Natural language processing. Thesis: *Efficient Parameter Adaptation Methods for LLMs: A Unifying Framework*. Supervisors: Prof. Mark Lee and Dr Phillip Smith.

MSc, Advanced Computer Science University of Birmingham 2018 – 2019

Distinction.

BEng, Computer Science Changsha University of Science and Technology 2013 – 2017

Outstanding Graduation Thesis award.

Technical skills

Modelling. PyTorch; fine-tuning with LoRA; pruning, low-rank compression and knowledge distillation; RLHF and DPO; multimodal understanding and generation; reinforcement learning and RL infrastructure.

Serving and applications. vLLM; LangChain, LangGraph and LangMem for multi-step workflows and assistants.

With the WGTECH team. Backend: Spring, Spring Boot and Spring Cloud. Web: Vue and Vite. iOS: Swift and Objective-C, MVC/MVVM, RxSwift/RxCocoa, Alamofire/Moya, Keychain and WCDB.

**Publications,
first author**

MoE-SVD: Structured Mixture-of-Experts LLMs Compression via Singular Value Decomposition

Wei Li, Lujun Li, Hao Gu, You-Liang Huang, Mark G. Lee, Shengjie Sun, Wei Xue, Yike Guo. *ICML 2025, PMLR 267:35209–35230*.

Compresses mixture-of-experts LLMs by decomposing expert layers into low-rank factors and sharing one factor across experts, with no additional training. The paper reports a 60% compression ratio and 1.5× faster inference with minimal performance loss on Mixtral, Phi-3.5, DeepSeek and Qwen2 models.

Adaptive Layer Sparsity for Large Language Models via Activation Correlation Assessment

Wei Li, Lujun Li, Mark Lee, Shengjie Sun. *NeurIPS 2024*.

Prunes layers by different amounts instead of applying one ratio everywhere, using correlations between layer activations to estimate how much each layer can lose. Evaluated on the LLaMA family and OPT, where it outperforms Wanda and SparseGPT, including at high sparsity.

BayesKD: Bayesian Knowledge Distillation for Compact LLMs in Constrained Fine-tuning Scenarios

Wei Li, Lujun Li, Mark G. Lee, Shengjie Sun, Lei Zhang, Wei Xue, Yike Guo. *Findings of ACL 2025, pp. 138–152*.

A distillation framework for recovering accuracy in compressed LLMs when the fine-tuning budget is limited. The paper reports average accuracy gains of 2.99% and 4.05% over standard distillation for 8B-parameter LLaMA and Qwen2 models.

**Publications,
co-author**

Cellular-Communication-Level Interpretability for Pathology Foundation Models via Graph Distillation on Microenvironment

Yuxiang Xiao, Zhiwei Chen, Dan Dai, **Wei Li**, Tianyang Zhang, Yakun Ju, Yang Hu, Kaixiang Yang. *iMIMIC Workshop at MICCAI 2026, accepted for oral presentation*.

Sub-MoE: Efficient Mixture-of-Expert LLMs Compression via Subspace Expert Merging

Lujun Li, Qiyuan Zhu, Jiacheng Wang, Xiaoyu Qin, **Wei Li**, Hao Gu, Sirui Han, Yike Guo. *AAAI 2026, 40(27):22994–23002*.

Outlier Matters: Efficient Long-to-Short Reasoning via Outlier-Guided Model Merging

Qiyuan Zhu, Dezhi Li, Lujun Li, Xiaoyu Qin, **Wei Li**, Hao Gu, Hua Xu, Sirui Han, Yike Guo. *AAAI 2026, 40(41):35213–35221*.

AIRA: Activation-Informed Low-Rank Adaptation for Large Models

Lujun Li, Dezhi Li, Cheng Lin, **Wei Li**, Wei Xue, Sirui Han, Yike Guo. *ICCV 2025, pp. 1729–1739*.

Efficient Fine-Tuning of Large Models via Nested Low-Rank Adaptation

Lujun Li, Cheng Lin, Dezhi Li, You-Liang Huang, **Wei Li**, Tianyu Wu, Jie Zou, Wei Xue, Sirui Han, Yike Guo. *ICCV 2025, pp. 22252–22262*.

Delta Decompression for MoE-based LLMs Compression

Hao Gu, **Wei Li**, Lujun Li, Qiyuan Zhu, Mark G. Lee, Shengjie Sun, Wei Xue, Yike Guo. *ICML 2025, PMLR 267:20497–20514*.

Academic service

Reviewer for *Machine Learning* (Springer), the NeurIPS 2025 Datasets and Benchmarks track, and ICLR 2026.

This is a selected list. The full record is on [OpenReview](#) and [Google Scholar](#).