

Mingxue (Mercy) Xu

Website | [X.com](#) | [Bluesky](#) | [LinkedIn](#) | [GitHub](#) | [Google Scholar](#) | [Email](#)

Education

- **PhD in Model Compression with Algebraic Approaches** Jan 2022 - May 2026
[Imperial College London](#), supervised by Prof. Danilo P. Mandic
Large Language Models (LLMs) compression with algebraic approaches.^{[1][2][3][5]}
Thesis: [Generative Language Model Compression with Algebraic Approaches](#).
- **MEng in Computer Science and Technology** Sept 2018 - July 2021
[University of Science and Technology of China](#), supervised by Prof. Xiang-Yang Li
Privacy leakage in federated learning systems^[9] and sensing for the Internet of Things (IoT)^[10].
Thesis: Set-level Membership Inference in Deep Learning Systems.
- **BEng in Computer Science and Technology** Sept 2014 - July 2018
[China University of Petroleum - Beijing](#), ranked 1st in the Dept.

Experience

- **Founder, Factthink LTD** May 2026 - present
 - Building [Graph-Hypo](#), a graph-harnessed multi-agent hypothesis generation system that develops research claims into candidate hypotheses and literature-grounded experiment plans.
 - Seven specialised agents extract concepts from literature, trace evidence, propose and critique hypotheses, then design experiments based on the built graph.
 - Combining deterministic vector-based approach and non-deterministic LLM-based verdicts.
 - Researchers' expertise and assigned topic dominate the whole pipeline.
- **PhD Researcher, Electrical and Electronic Engineering Dept. at Imperial** Jan 2022 - May 2026
 - Generative language compression with algebraic approaches.
 - ◇ Used representation dimension to estimate LLM performance after structural pruning.^[2]
 - ◇ Unified terminology of matrices and tensors in the context of generative language models^[3], for the interpretation of generative language model efficiency.
 - ◇ Compressed the GPT series models based on tensor decomposition and deployed the compressed models on low-end devices (i.e. Raspberry Pi).^[1]
 - Converted Membership Inference into an approach of dataset legitimacy assessment (i.e. fairness and diversity) in the setting of Machine Learning as a Service^[6], which was featured by [Montreal AI Ethics Institute](#).
- **LINKE Lab, University of Science and Technology of China (MEng)** Apr 2018 - Jun 2021
 - Implemented dataset privacy (subject property information) inference in the Federated Learning setting with CycleGAN^[9].
 - Implemented lightweight LSTM models on smartwatches for user identification & authentication^[10] with the gesture signal collected by accelerometer and gyroscope.
- **Baidu PaddlePaddle, Technical Product Manager (intern)** Oct 2021 - Dec 2021
 - Co-authored a paper on the product design of [PaddlePaddle](#), a deep learning platform for users of various backgrounds (i.e. technical or non-technical)^[8].
 - Served for the developer communities of [PaddleSpeech](#).

Selected Publications & Manuscripts

- [1] **Mingxue Xu**, Yao Lei Xu and Danilo P. Mandic. (2026). [“Surviving Resource-constraint Compression: Capability Retention under Tensor-train Decomposition for Sub-billion Parameter Language Models”](#), *Machine Intelligence Research*, 23(3), Springer. Extended from [“TensorSLM”](#), *ICML 2025 Workshop on Tiny Titans (TTODLer-FM)*
- [2] **Mingxue Xu**, Lisa Alazraki and Danilo P. Mandic. (2025). [“How can representation dimension dominate structurally pruned LLMs”](#), *ICLR 2025 Workshop on Sparsity in LLMs (SLLM)*
- [3] **Mingxue Xu**, Sadia Sharmin and Danilo P. Mandic. (2024). [“Geometry is All You Need: A Unified Taxonomy of Matrix and Tensor Factorization for Compression of Generative Language Models”](#), *arXiv*
- [4] Giorgos Iacovides, Thanos Konstantinidis, **Mingxue Xu**, Danilo P. Mandic. (2024). [FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading](#), *International Conference on AI in Finance (ICAIF2024)*
- [5] **Mingxue Xu**, Yao Lei Xu and Danilo P. Mandic. (2023) [“TensorGPT: Efficient Compression of the Embedding Layer in LLMs based on the Tensor-Train Decomposition”](#), *arXiv*
- [6] **Mingxue Xu**, Tongtong Xu and Po-Yu Chen. (2023). [“Private Training Set Inspection in MLaaS”](#), *arXiv*
- [7] **Mingxue Xu** and Xiang-Yang Li. (2022). [“Data Origin Inference in Machine Learning”](#), *arXiv*
- [8] Ran Bi, Tongtong Xu, **Mingxue Xu** and Enhong Chen. (2022). [“PaddlePaddle: A Production Oriented Deep Learning Platform Facilitating the Competency of Enterprises”](#), *24th IEEE International Conference on High Performance Computing & Communications, 2022*
- [9] **Mingxue Xu** and Xiang-Yang Li. (2020). [“Subject Property Inference Attack in Collaborative Learning”](#), *12th IEEE International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC2020)*
- [10] Xiaojing Yu, Zhijun Zhou, **Mingxue Xu**, Xuanke You and Xiang-Yang Li. (2020). [“ThumbUP: Identification and Authentication by Smartwatch using Simple Hand Gestures”](#), *18th Annual IEEE International Conference on Pervasive Computing and Communications (PerCom2020)*

Open Source Projects

- **llm-profiler**: a fine-grained LLM profiler that calculates the algorithmic operations (e.g. ADD, MULT), parameters, cache (buffers) and activations (intermediate output between the layers).
- **omnipies**: turning vague research ideas and questions into structured, evidence-backed arguments, based on the architecture proposed in *The Craft of Research*, by Booth et al.
- **graph-hypoth**: a multi-agent system that balances creativity and verifiability (auditability), centering a **graph** to together work on.

Awards and Honors

- | | |
|--|------------------|
| • First Prize Scholarship in University of Science and Technology of China | 2019, 2020, 2021 |
| • Outstanding Graduate Student of Beijing | 2018 |
| • Meritorious Winner in Mathematical Contest in Modeling (top 10%), U.S. | 2016 |
| • National Scholarship (ranked 1 st in the department) | 2016, 2015 |

Technical Skills

- **Programming**: proficient in Python and C/C++.
- **Systems**: profiled ML models for memory usage, latency and FLOPs (see **llm-profiler**); served as GPU (Ubuntu) server administrator of PhD research group; experience with deploying distributed edge computing system with Docker and Kubernetes.