

XU YUANJIAN(徐元健)

Tencent | MSRA | THU | HKUST | PKU | Nankai

phone: (+86) 17637303503 · email: xuyuanjian233@gmail.com · [homepage](#)

EDUCATION

 Hong Kong University of Science and Technology , Ph.D. in Fintech	2023.09–2027.04
 Peking University , Master in Computer Science	2020.09–2023.06
 Nankai University , Bachelor in Computer Science	2014.09–2018.06

A BRIEF INTRODUCTION

I am a Ph.D. candidate at the Hong Kong University of Science and Technology, Guangzhou campus. I am supervised by Prof. Guang Zhang from HKUST-GZ. My research interest revolves around **Data-centric Machine Learning (LLMs)**. Specifically, my work has systematically investigated the following four dimensions:

- **RQ1 (Data Selection):** Identifying optimal data selection strategies for full-lifecycle LLM training.
- **RQ2 (Data Curriculum):** Designing effective data scheduling to maximize model performance.
- **RQ3 (Data Representation):** How to model complex data characteristics (e.g. multi-modal data)?
- **RQ4 (Data Orchestration):** How to elegantly organize heterogeneous data (e.g. graph)?

To answer these questions, I have published several papers at leading venues as the first author including **ACL 2023**, **ACL 2025**, **ACL 2026** (×2), **ICML 2026** (×2), **ACM ICAIF 2025**, and **ICASSP 2026**. I also serve as a reviewer for leading conferences such as **NeurIPS**, **ICLR**, and **ICML** ( **Gold Reviewer Award**). At the same time, I was selected for top talent internship programs at **Tencent HunYuan** and **Tsinghua AIR**.

SELECTED PUBLICATIONS (FIRST AUTHOR)

- [**ICML 2026 (Spotlight)** ] **Yuanjian Xu**, et al. “Towards Efficient LLMs Annealing with Principled Sample Selection.” [CCF A] [CORE A*] @ MSRA 
- [**ICML 2026**] **Yuanjian Xu**, et al. “D³: Dynamic Directional Graph-Constrained Data Scheduling for LLM Training.” [CCF A] [CORE A*] @ MSRA 
- [**ACL 2026**] **Yuanjian Xu**, et al. “DoGraph: Towards Domain-Aware LLM Training with Graph-Guided Weighting.” [CCF A] [CORE A*]
- [**ACL 2026**] **Yuanjian Xu**, et al. “BizCompass: Benchmarking the Reasoning Capabilities of LLMs in Business Knowledge and Applications.” [CCF A] [CORE A*]
- [**ACL 2025**] **Yuanjian Xu**, et al. “FinRipple: Aligning Large Language Models with Financial Market for Event Ripple Effect Awareness.” [CCF A] [CORE A*]
- [**ACL 2023**] **Yuanjian Xu**, et al. “Hard Sample Aware Prompt-Tuning.” [CCF A] [CORE A*]
- [**ICASSP 2026 (Oral)** ] **Yuanjian Xu**, et al. “Hermite Discriminator for Neural SDEs.” [CCF B] [CORE A]
- [**ICAIF 2025**] **Yuanjian Xu**, et al. “LENS: Large Pre-trained Transformer for Exploring Financial Time Series Regularities.” (Leading conference for AI in finance)

ONGOING WORKS (FIRST AUTHOR)

- [**Target ICLR 2027**] **Yuanjian Xu**, et al. “Identifying Knowledge Gaps via the Manifold Geometry of Soft Prompts.” [CCF A] @ Tencent Hunyuan 
- [**Target ICLR 2027**] **Yuanjian Xu**, et al. “Benchmarking Sample Reasoning Quality from Internal Geometric Dynamics in LLMs.” [CCF A] @ Tencent Hunyuan 
- [**ACL ARR (Under Review)**] **Yuanjian Xu**, et al. “The WoW: A Large-Scale World Knowledge Corpus for Full-Lifecycle LLM Training.” [CCF A]

- [**ACL ARR (Under Review)**] **Yuanjian Xu**, et al. “From Tokens to Intelligence: A Survey on Data Selection for Large Language Models.”

INTERNSHIP EXPERIENCE

 **Tencent HunYuan** QingYun Program Research Intern | 2026.5–Present
Focused on reasoning data quality assessment and zero-annotation evaluation for large language models.

 **Joinquant** (Billion-scale quantitative fund) Research Intern | 2026.02–2026.05
Worked under the supervision of PM Ruixiao. Developed tick-level generative and representation models for high-frequency trading. Addressed key challenges including non-equally spaced data and market randomness.

 **Microsoft Research Asia (MSRA)** Tomorrow Star Research Intern | 2025.10–2026.02
Worked under the supervision of Dr. Zhong Li. Focused on data selection and training order optimization for large language models. Proposed the D³ and DiReCT methods, both accepted at ICML 2026.

 **AIR, Tsinghua University** Research Intern | 2022.03–2023.03
Worked under the supervision of Prof. Zaiqing Nie. Contributed to the Meituan Nutrition Knowledge Graph construction. Investigated hard sample problems in NLP and proposed HardPT, published at ACL 2023.

ACADEMIC SERVICE

Reviewer: ICML (2025, 2026), NeurIPS (2024, 2025, 2026), ICLR (2025, 2026), AAAI 2026, COLM 2026

Hong Kong University of Science and Technology, Teaching Assistant, Advanced Statistics

SKILLS

Machine Learning: I am well-versed in modern generative models, LLMs, and machine learning.
Mathematical Foundation: I have a relative solid mathematical foundation. During my Ph.D. studies, I passed the qualifying exam in optimization, advanced probability theory, advanced statistics, and stochastic analysis.
Financial Theory: I also have exposure to financial theory, having passed FRM Level 1 and studied advanced microeconomics and advanced asset pricing.

HONORS & AWARDS

-  Award for Excellent Academic Excellence, Peking University 2021 Certificate No.: H2021000170320
-  Full Ph.D. Scholarship, Hong Kong University of Science and Technology 2023–Present
-  Tomorrow Star Plan (Top 10% Internship Program), Microsoft Research Asia 2025
-  QingYun Program (Top Talent Internship Program), Tencent Hunyuan 2026
-  ICML Gold Reviewer Award 2026