

Meihao Fan

+86-198-0038-5890 | fmh1art@ruc.edu.cn | [fmh1art.github.io](https://github.com/fmh1art)

 Meihao Fan |  fmh1art |  fmh1art

Haidian District, Beijing 100000, China

SUMMARY

Ph.D. student in Computer Science at Renmin University of China, focusing on **LLM-powered data agents**, **reinforcement learning**, and **post-training for open-source LLMs**. My research centers on building **cost-effective and reliable agentic systems** for data-intensive tasks such as data preparation, entity resolution, and autonomous data science. More broadly, I am interested in developing lightweight agents with stronger **data awareness** to optimize deep research workflows, reduce expensive trial-and-error, and improve the practical efficiency of LLM-based systems.

EDUCATION

- **Renmin University of China** Sept 2023 - Expected June 2028
Ph.D. in Computer Application Technology Beijing, China
 - **Honors:** National Scholarship (The **only** 2nd-year Ph.D. recipient in the college)
 - **Advisor:** Prof. Ju Fan / Prof. Xiaoyong Du
- **Chongqing Jiaotong University** Sept 2019 - June 2023
B.S. in Computer Science and Technology Chongqing, China
 - **GPA:** 4.31/5.00
 - **Honors:** National Scholarship (The **only** recipient in the college), Mingde Scholarship Nomination Award (Top 20 in the university)

EXPERIENCE

- **ByteDance**  March 2025 - March 2026
Research Intern (LLM Data Agent Systems) Beijing, China
 - Led the research and development of **DeepPrep**, an **LLM-powered data agent** system for autonomous data preparation, targeting reliable and cost-effective execution in data-intensive workflows.
 - Proposed a **tree-based agentic reasoning** mechanism with execution-grounded feedback and non-local backtracking, improving the **reliability** and **steerability** of data agents.
 - Designed a **progressive post-training** framework for open-source LLMs, combining cold-start supervised fine-tuning and multi-turn **reinforcement learning** with a hybrid reward.
 - Built large-scale benchmarks via reversible noise injection and demonstrated that the system achieves strong performance at up to **15× lower inference cost**, highlighting its value for **cost-efficient agent systems**.
- **Renmin University of China** Sept 2023 - Present
Research Assistant Beijing, China
 - Led research on **AutoPrep**, a multi-agent data preparation framework that maps natural language analytical requests into executable data workflows.
 - Studied **cost-effective LLM inference** for entity resolution and proposed the **BATCHER** framework, balancing task quality and API cost in in-context learning.
 - Co-authored research on **autonomous data agents**, **text-to-SQL reasoning**, and data-centric LLM systems.

PROJECTS

- **DeepPrep: Data Agent System for Autonomous Data Preparation** March 2025 - March 2026
Data Agents, Tree-based Reasoning, Post-Training, RL, Qwen 
 - Developed an execution-grounded environment supporting 31 data preparation operators, enabling agents to materialize intermediate table states and leverage runtime feedback for decision making.
 - Implemented a full **post-training** pipeline for open-source LLMs, covering operator syntax learning, supervised fine-tuning, and multi-turn **reinforcement learning**, with trained model weights released from 0.5B to 14B.
- **AutoPrep: Multi-Agent Data Preparation Framework** 2024 - 2025
LLMs, Data Agents, Multi-Agent Systems 
 - Designed a planner agent with a Chain-of-Clauses reasoning mechanism for high-level workflow planning.
 - Developed programmer agents with tool-augmented code generation to translate natural language requests into executable data operations.
- **BATCHER: Cost-Effective LLM Inference for Entity Resolution** 2023 - 2024
In-Context Learning, Prompt Engineering, Cost Optimization, Batch Prompting 
 - Introduced a batch prompting framework with demonstration selection and question batching for entity resolution.
 - Proposed a covering-based demonstration selection strategy, achieving **4x-7x cost savings** over standard prompting while maintaining strong matching quality.

- [C.1] **Meihao Fan**, Ju Fan, Yuxin Zhang, Shaolei Zhang, Xiaoyong Du, Jie Song, Peng Li, Fuxin Jiang, Tiejing Zhang, Jianjun Chen. DeepPrep: An LLM-Powered Agentic System for Autonomous Data Preparation. VLDB 2026 [Under Review]
- [C.2] **Meihao Fan**, Ju Fan, Nan Tang, Lei Cao, Guoliang Li, Xiaoyong Du. AutoPrep: Natural Language Question-Aware Data Preparation with a Multi-Agent Framework. VLDB 2025 [Accepted]
- [C.3] **Meihao Fan**, Xiaoyue Han, Ju Fan, Chengliang Chai, Nan Tang, Guoliang Li, Xiaoyong Du. Cost-Effective In-Context Learning for Entity Resolution: A Design Space Exploration. ICDE 2024 [Accepted]
- [C.4] Yuxin Zhang, **Meihao Fan**, Ju Fan, Mingyang Yi, Yuyu Luo, Jian Tan, Guoliang Li. Reward-SQL: Boosting Text-to-SQL via Stepwise Execution-Aware Reasoning and Process-Supervised Rewards. SIGMOD 2026 [Accepted]
- [C.5] Chao Deng, Ju Fan, Yuyu Luo, Qinliang Xue, **Meihao Fan**, Yuxin Zhang, Min Zhang, Xiaofeng Jia, Jing Zhang, Xiaoyong Du. TACO: A Benchmark for Open-Domain Text-to-SQL with Ambiguous and Cross-Database Queries. VLDB 2026 [Accepted]
- [C.6] Anonymous Authors (incl. **Meihao Fan**). DeepAnalyze: Agentic Large Language Models for Autonomous Data Science. ICML 2026 [Under Review]
- [C.7] Anonymous Authors (incl. **Meihao Fan**). CODA-BENCH: Can Code Agents Handle Data-Intensive Tasks? ICML 2026 [Under Review]

HONORS AND AWARDS

- **National Scholarship (Ph.D.)** 2025
Ministry of Education, China
 - Awarded to top-performing graduate students nationwide.
 - The **only** second-year Ph.D. student in the college to receive this prestigious award.
- **National Scholarship (B.S.)** 2023
Ministry of Education, China
 - The **only** recipient in the college.
- **Mingde Scholarship Nomination Award** 2023
Chongqing Jiaotong University
 - Awarded to the top 20 students in the entire university.

SKILLS & RESEARCH INTERESTS

- **Research Interests:** Data Agents, Reinforcement Learning for LLM Agents, Post-Training, Cost-Efficient LLM Systems, Autonomous Data Science, Text-to-SQL
- **Programming & Tools:** Python, LaTeX
- **LLM & Agent Systems:** Agentic Reasoning, Multi-Agent Systems, Supervised Fine-Tuning (SFT), Multi-turn RL, Prompt Engineering, Execution-Grounded Feedback
- **Data Systems:** Data Preparation, Entity Resolution, Data Integration, Analytical Workflows