Zhenyu Wu

PhD Data Scientist | Structured Data, Knowledge Graphs & AI Agents

Manchester, UK | +44 7741541914 | github.com/PierreWoL

Summary

- PhD Data Scientist with four years of research experience developing machine-learning systems for heterogeneous data repositories, with emphasis on schema, taxonomy, ontology, and knowledge inference.
- Built open-source Python/PyTorch pipelines spanning Transformer representation learning, contrastive learning, decoder-only LLM inference, hierarchical clustering, robustness analysis, and reproducible evaluation.
- Research published in or accepted by leading peer-reviewed venues, including *The VLDB Journal*, SIGMOD, and ESWC; experienced in communicating technical methods through international conference presentations and an EDBT tutorial.
- Developing agentic data-analysis methods that ground multi-database question answering and Text-to-SQL workflows in reliable schema linking and structured outputs.

Education

University of Manchester Manchester, UK

Ph.D. candidate, Department of Computer Science Sep 2022 – Expected Oct 2026

- Research: LLM-based structured-data understanding, schema/taxonomy inference, knowledge graphs, and agentic data analytics
- Supervisors: Prof. Norman Paton and Dr. Jiaoyan Chen; Dean's Doctoral Scholarship (one of three departmental awards)

University of Edinburgh – *M.Sc. High Performance Computing with Data Science (Merit)* Sep 2019 – Aug 2020

Tianjin University of Finance and Economics – *B.Sc. Information Management & Information Systems* Sep 2015 – Jun 2019

- GPA: 3.81/4.00 (Top 5%); Minor in Accounting

Research Experience

LLM-Based Schema and Knowledge Inference for Heterogeneous Data Repositories

Built modular research systems that transform raw table headers and values into machine-readable schemas and knowledge structures. Code: github.com/PierreWoL/SILM.

- Designed **EmSI**, combining Transformer representation learning with hierarchical clustering, and **GeSI**, using decoder-only LLM inference with batched/cached execution; recovered entity types, attributes, hierarchies, and cross-type relationships.
- Established reproducible evaluation and controlled robustness studies across four real-world datasets, varying schema size, attribute overlap, and header noise; exceeded SOTA baselines by 30–40% in F1/Recall and 7–10% in Rand Index.

Contrastive Representation Learning and Ontology Construction for Table Repositories

Developed encoder- and decoder-based systems for inducing entity-type taxonomies and supporting semantic linking across heterogeneous data. Code: github.com/PierreWoL/TwoMethods.

- Adapted SwAV-style in-cluster contrastive learning to align augmented and semantically related columns, improving Rand Index by 10% and Purity by 18%.
- Compared **EmTT** (encoder-based hierarchical clustering) with **GeTT** (decoder-based top-down Chain-of-Layer inference) across three datasets; achieved 0.85–1.0 tree consistency and above 0.9 top-level RI/Purity, with ablations documenting accuracy–stability trade-offs.

Additional Applied Research

Agentic Multi-Database Question Answering and Schema Linking *(Ongoing)*

- Developing an LLM-based schema-linking component for multi-database question answering and downstream Text-to-SQL/data-analysis agents.
- Built a Qwen2.5-14B pipeline combining question-to-attribute matching with query-time augmentation, improving schema-matching accuracy by 20%; designed structured database/table/attribute outputs and failure analysis for ambiguous cross-database queries.

Visiting Research Associate, Singapore University of Technology and Design Apr 2021 – Mar 2022

- Co-built a unified benchmark for data-stream clustering on real and synthetic streams, comparing algorithms across summary structures, window models, outlier handling, and offline refinement; work accepted by SIGMOD 2023.
- Implemented four representative algorithms in a shared codebase with consistent interfaces/configuration, end-to-end runners, and sanity checks for reproducible large-scale evaluation. Code: github.com/intellistream/Sesame.

Selected Publications

- **Zhenyu Wu**, Jiaoyan Chen, Norman W. Paton. *Schema Inference for Tabular Data Repositories Using Large Language Models*. Accepted by *The VLDB Journal*.
- **Zhenyu Wu**, Jiaoyan Chen, Norman W. Paton. *Taxonomy Inference for Tabular Data Using Large Language Models*. *ESWC*, 2025.
- Norman W. Paton, Jiaoyan Chen, **Zhenyu Wu**. *Dataset Discovery and Exploration: A Survey*. *ACM Computing Surveys*, 2024.
- Xin Wang, Zhengru Wang, **Zhenyu Wu**, et al. *Data Stream Clustering: An In-depth Empirical Study*. *SIGMOD*, 2023.

Communication & Recognition

- Delivered an oral research presentation at ESWC 2025 and co-presented an EDBT 2024 tutorial on dataset discovery, translating methods, evidence, and limitations for international audiences.
- Dean's Doctoral Scholarship, University of Manchester (one of three departmental awards), 2022.

Technical Skills

Programming & Data: Python, SQL, C++, Linux/Shell, Git
ML/LLM: PyTorch, Hugging Face Transformers, self-supervised and contrastive learning, hierarchical clustering, vLLM
Knowledge & Agents: schema/taxonomy/ontology inference, knowledge graph construction, LangChain, LangGraph, tool calling, structured validation
Engineering & Evaluation: modular Python pipelines, interface-based design, experimental design, statistical analysis, robustness/ablation studies, reproducible benchmarking
Languages: English, Mandarin