

Qiaoyuan Zheng

📧 qiaoyuan667 | 📞 +41 76 625 10 53 | ✉️ zqiaoyuan@ethz.ch | 🌐 qiaoyuan-zheng |
💻 qiaoyuan-zheng.com | 📍 Zurich, Switzerland | 🎓 MSc in Data Science



MSc student in Data Science at ETH Zürich, researching reliable LLM evaluation, interpretable capability assessment, and privacy in language agents.

Education

ETHzürich ETH Zürich

Zurich, Switzerland

M.Sc. DATA SCIENCE

2025 - Present

- GPA: 5/6
- Course: Probabilistic AI | Large Language Models | Big Data | Computational Statistics | Large-Scale AI Engineering
- Special Project-oriented course: AI Center Project | Research in DS | Data Science Lab

Sorbonne University

Paris, France

B.Sc. COMPUTER SCIENCE/MATHEMATICS

2022 - 2025

- Math Course: Linear Algebra | Digital and Functional Series | Topology | Theory of Measurement | Theory of Probability
- CS Course: Algorithms | Data Structure | Object-oriented Programming | Functional Programming | Advanced C
- GPA: 17.10/20, 3.96/4.0 (WES), L3 Mention: Bien

Research Experience

ETH AI Center

Zurich, Switzerland

RESEARCH ASSISTANT

2026 - Present

- Develop a multilingual benchmark for evaluating legal compliance in large language models
- Building reproducible pipelines for scenario generation, independent review, and model evaluation
- Collaborate with legal experts to assess scenario plausibility and legal validity
- Investigate benchmark applications for LLM post-training

Apertus LLM Evaluation Python

ESIEE Paris

Paris, France

2023 - 2023

RESEARCH ASSISTANT · GITHUB

- Implemented Transformer-based models and PatchTST for time-series forecasting under the supervision of Dr. You Jiang
- Adapted data and training pipelines, comparing model performance and computational efficiency against U-Net 24 on synthetic datasets

Python PyTorch Time-Series Forecasting

Publications

POLAR-Bench: A Diagnostic Benchmark for Privacy-Utility Trade-offs in LLM Agents

2026

Qiaoyuan Zheng, Yiqu Yang, Qi Gao, Imanol Schlag

preprint, Under Review at NeurIPS 2026 E&D Track

- Introduced a diagnostic benchmark covering 7,852 samples across 10 domains to evaluate how LLM agents balance task utility with user-defined privacy constraints.
- Separated privacy-policy dimensions and adversarial strategies in a 5×5 evaluation design, using deterministic scoring to identify specific failure conditions.
- Identified substantial differences in privacy protection between frontier models and smaller open-weight models, with the weakest evaluated models leaking over half of protected attributes.

Are Near-Tied LLM Rankings Robust to Family-DIF-Guided Benchmark Recomposition?

2026

Qiaoyuan Zheng, Yiqu Yang

Preprint, Under review at ICLR 2027 and the NeurIPS TAE Workshop

- Investigated the robustness of near-tied LLM rankings across five benchmarks using a spectral approximation to multidimensional item response theory and differential item functioning analysis.
- Used owner-disjoint evaluation folds and matched-random controls to distinguish systematic effects of benchmark composition from generic subtest variation.

SkillEval: Learning Interpretable Ability Profiles of LLMs via Cognitive Diagnosis Models

2026

Berke Arda, Peng Cui, **Qiaoyuan ZHENG**, Rudolf Debelak, Mubashara Akhtar, Mrinmaya Sachan

[Under Review at NeurIPS 2026 E&D Track](#)

- Proposed a framework combining LLM-based skill discovery with neural cognitive diagnosis to construct interpretable ability profiles for 3,811 LLMs across 9,523 items from five benchmarks.

Last Translation Benchmark

2026

Vilém Zouhar, Niyati Bafna, Mukund Choudhary, ..., **Qiaoyuan Zheng**, ...

[Preprint](#)

- Contributed translation test cases and reviewed English, Chinese, and French benchmark submissions.

Selected Projects

Sorbonne University

APERIODIC TILING GENERATION

Mar 2024 - Jul 2024

- Implemented Penrose, Wang, and Hat tiling algorithms in Python.
- Optimized generation through duplicate detection and spatial pruning; analyzed runtime scaling.

Python

Algorithms

ESIEE Paris

Q-LEARNING AND SARSA IN GRID WORLDS

Jul 2023 - Aug 2023

- Implemented and compared Q-learning and SARSA across 15 custom grid-world environments.
- Built a map editor and environments adopted for master's-level teaching at UPEC.

Python

Reinforcement Learning

4EU+ Alliance Research Program

COVID-19 MOBILITY ANALYSIS

Feb 2020 - Dec 2021

- Led a four-person team studying cross-country mobility patterns during COVID-19.
- Used linear regression and visualization to examine associations between public-health policies and time spent at home.

Python

Statistical Analysis

Skills

Languages

●●●●○ Chinese (Native) | English (Fluent, IELTS 7.5, GRE 328) | French (Fluent, Delf B2)

Programming

●●●●● Python | C | Java | R | Ocaml | SQL

Tech Stack

Docker | Git

Frameworks & Lib

Pandas | NumPy | Scikit-learn | Pytorch

Personal Interests

Musical instruments (Drums, Saxophone, etc.) | Cooking | Reading | Calligraphy