Yiwei Liu

Lausanne, Switzerland | yiw.liu@epfl.ch | ivyyyyw.github.io | Yiwei Liu | github

Education

École Polytechnique Fédérale de Lausanne (EPFL) — Lausanne, Switzerland
MSc in Digital Humanities (GPA: 5.3/6) — *Sept* 2023 – *Now*
- Relevant Coursework: Modern Natural Language Processing, Applied Data Analysis, Computational Social Media, Reinforcement Learning, Large-scale Data Science for Real-world Data.

ETH Zürich — Zürich, Switzerland
Semester Project Student (GPA: 5.75/6) — *June* 2024 – *Dec* 2024

Beijing Institute of Technology (BIT) — Beijing, China
BSc in Statistics (GPA: 88/100) — *Sept* 2019 – *June* 2023
- Relevant Coursework: Optimization, Machine Learning, Matrix Analysis, Function of Real Variables, Multivariate Statistical Analysis, Discrete Mathematics, Analysis.

Research Interest

I'm interested in evaluating and aligning AI systems through the lens of human cognition (linguistics, psychology, social reasoning, etc.) to build safer AI systems that align with human values and behaviors.

Publications

Conference Publication

- **TactfulToM: Do LLMs Have the Theory of Mind Ability to Understand White Lies?**
  Author: Yiwei Liu, Emma Jane Pretty, Jiahao Huang, Saku Sugawara
  Accepted at *EMNLP 2025* (main)

Book Chapter

- **Business Model Construction Logic of Research-based Platforms**
  Co-author of Chapter 12 on Biogen's collaborative R&D model
  Published by China Machine Press, 2024 (ISBN: 978-7-111-73714-8)

Research Experience

Student Researcher — Tokyo, Japan
National Institute of Informatics Supervisor: Prof. Saku Sugawara — *Oct* 2024 – *Mar* 2025; *May* 2025 – *now*
- Designed TactfulToM benchmark (100 samples, 6.7k questions) to assess LLMs' Theory of Mind reasoning in white lie contexts; evaluated 9 state-of-the-art models and revealed significant model-human performance gaps, highlighting critical alignment challenges.
- Evaluating LLMs' pragmatic reasoning and common ground understanding in ambiguous contexts; designing alignment methods to improve clarification-seeking behavior for more efficient human-AI communication. *[Ongoing—preparing for ACL 2026 submission]*

Semester Project Student — Zurich, Switzerland
Doctor Advisor: Robin Chan | Supervisor: Prof. Menna El-Assady, Prof. Frédéric Kaplan — *June* 2024 – *Dec* 2024
- A Visual Resolution for Syntactic Ambiguity: developed a k-best constituency parser and an interactive web interface, enabling detection and exploration of different kinds of syntactic ambiguity via comparison of alternative parses and complementary visualizations.
- Evaluated how LLMs process ambiguity to identify alignment gaps with human interpretation.

Internship Experience

AXA GO Research — Lausanne, Switzerland
Research Intern | Supervisor: Dr. Thibault Laugel — *Aug* 2025 – *Now*
- Evaluate knowledge editing methods (locate-then-edit, fine-tuning-based, etc.) to identify performance boundaries across diverse models and tasks. Apply interpretability tools like PatchScopes to analyze how test-time intervention propagate or fail, revealing inherent limitations of these methods.

Beijing Xianfeng Changqing Venture Capital (K2VC) Beijing, China

Investment Research Intern | Advisor: Runxin Yang *Dec* 2022 – *June* 2023

- Conducted in-depthin dustry research on LLMs and their applications, engaging with startups to evaluate innovations; authoring reports on advancements and emerging trends in the LLM ecosystem supporting investment decision.

MiraclePlus (Former Y Combinator China) Beijing, China

Investment Research Intern |Advisor: Jack Jin, Dr. Qi Lu *Dec* 2021 – *Feb* 2022

- Engaged with leading AI labs in China to understand frontier research directions; performed technical due diligence on investment targets including TensorOpera AI and Thewake Systems, leading to successful investments.

JD.com Explore Academy (AI Lab) Beijing, China

Algorithm Engineer Intern| Advisor: Dr. Chang Li *June* 2021 – *Aug* 2021

- Built Proof-of-Concept for JD.com × Nvidia Omniverse Virtual Retailer that collects eye-tracking and interaction data and applies machine learning to analyze data for shelf layout optimization.
- Co-authored AI research strategy for JD Explore Academy, focusing on multimodal and trustworthy AI for e-commerce.

Selected Projects

Developed a specialized LLMs tutor for EPFL course *Feb* 2024 - *May* 2024

Course project for Modern Natural Language Processing | Instructor: Antoine Bosselut

- Fine-tuned the Phi-2 model with DoRA (PEFT) and Direct Preference Optimization (DPO) for efficient alignment; integrated Retrieval-Augmented Generation (RAG) for context-aware responses and applied quantization to optimize the deployment.

Advantage Actor-Critic (A2C) for the CartPole problem *Feb* 2024 - *May* 2024

Course project for Reinforcement Learning | Instructor: Gerstner Wulfram

- Implemented the Advantage Actor-Critic (A2C) algorithm on the CartPole environment, leveraging parallel processing and multi-step returns to improve training efficiency.

Extending Text2Image Model to Multi-Modal Model *Sept* 2023 - *Dec* 2023

Course project for Foundation of Digital Humanity | Instructor: Antoine Bosselut

- Fine-tuned the Phi-2 model using DoRA (PEFT) and Direct Preference Optimization (DPO) for efficient alignment; integrated Retrieval-Augmented Generation (RAG) for context-aware responses and applied quantization to optimize deployment.

Beer Review Data Analysis *Sept* 2023 - *Dec* 2023

Course project for Applied Data Analysis (Best 10 projects) | Instructor: Robert West

- Applied Sentence-BERT to analyze beer reviews from RateBeer and BeerAdvocate, generating emotion scores and building a mood-based beer recommendation system to support marketing strategies.

Honors and Awards

- Awarded "Gintong Talent", Renmin University of China — 35 students selected annually 2022
- Awarded Academic Excellent Scholarship, Beijing Institute of Technology — Top 20% 2020–2021, 2021–2022

Service

- Served as mentor and in the event coordination team for "Forward With Her", a mentorship program dedicated to empowering women in STEM. *Sept* 2024 – *Now*

Skills

- **Programming & Frameworks:** Python, PyTorch, TensorFlow, Sklearn, Pandas, NumPy, JavaScript/HTML, Git, SQL, Linux, LaTeX
- **Domains:** AI, Machine Learning, Cognition, Ethics
- **Languages:** English (Proficient), Chinese (Native), French (Beginner), German (Beginner)