Yiwei Liu

Lausanne, Switzerland | yiw.liu@epfl.ch | (+41) 76 824 92 89 | Yiwei Liu

EDUCATION

École Polytechnique Fédérale de Lausanne (EPFL) — Lausanne, Switzerland
MSc in Digital Humanities (GPA: 5.3/6) — *Sept 2023 - Now*

Beijing Institute Of Technology (BIT) — Beijing, China
BSc in Statistics (GPA: 88/100) — *Sept 2019 - June 2023*

PUBLICATION

TactfulToM: Do LLMs Have The Theory Of Mind Ability To Understand White Lies *link*
Yiwei Liu, Emma Jane Pretty, Jiahao Huang, Saku Sugawara
Accepted at EMNLP Main 2025. (Conference Paper)

Business Model Construction Logic Of Research-Based Platforms
Co-author of Chapter 12 on Biogen's collaborative R&D model, with contributions in data extraction and analysis.
China Machine Press, ISBN 978-7-111-73714-8 (Book Chapter)

RESEARCH EXPERIENCE

Evaluating LLMs' Social Reasoning Abilities On White Lies — National Institute of Informatics, Japan
Research Assistant | Supervisor: Saku Sugawara — *Oct 2024 – Now*

- Designed and developed a novel benchmark to evaluate large language models' (LLMs) Theory of Mind ability to detect and reason about white lies in real-life conversations, and tested on 15+ the state-of-art models.

A Visual Resolution For Syntactic Ambiguity *link* — ETH Zürich, Switzerland
Research Student | Doctor Advisor: Robin Chan | Supervisor: Menna El-Assady, Frédéric Kaplan — *July 2024 – Nov 2024*

- Developed a K-best constituency parsing-based model to detect syntactic ambiguity, enabling identification of multiple valid syntactic interpretations.
- Built an interactive visualization interface, and designed multiple methods to explore alternative interpretations of syntactically ambiguous sentences.
- Evaluated the ability of LLMs to recognize and differentiate ambiguous structures.

Analyzing Digital Governance Measures For COVID-19 Tracing Policies — Renmin University, China
Research Assistant | Supervisor: Peng Wang — *Sept 2021 – Nov 2021*

- Conducted data analysis on the effectiveness of digital tracing technologies (e.g., QR codes) in controlling the spread of COVID-19, by collecting and preprocessing digital governance data, and applying Python for statistical analysis and visualization.

INTERNSHIP EXPERIENCE

AXA Group Operations — Lausanne, Switzerland
Research Intern | Advisor: Thibault Laugel — *Aug 2025 - Now*

- Research on Knowledge Editing and Activation Steering for LLM Alignment.

Beijing Xianfeng Changqing Venture Capital (K2VC) — Beijing, China
Investment Research Assistant | Advisor: Runxin Yang — *Jan 2023 – June 2023*

- Conducted in-depthin dustry research on LLMs and their applications, engaging with startups to evaluate technological innovations and assess market potential.

MiraclePlus (Former Y Combinator China) — Beijing, China
Investment Research Assistant | Advisor: Jack Jin, Dr. Qi Lu — *Dec 2021 – Feb 2022*

- Conducted research on ML technologies, targeting frontier AI Labs in China.
- Performed comprehensive due diligence on investment targets such as TensorOpera AI and Thewake Systems, involving technical evaluation and entrepreneur interviews, which led to successful investment decisions.

JD.Com Explore Academy (AI Lab) — Beijing, China
Algorithm Engineer | Advisor: Dr. Chang Li — *June 2021 – Aug 2021*

- Participated in building a proof-of-concept for the JD × Nvidia Omniverse Virtual Retailer project, showcasing the collection of eye-tracking and interaction data between customers and the environment; co-authored an AI research strategy plan for JD Explore Academy, focusing on multimodal technologies and trustworthy AI

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 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.

Advantage Actor-Critic (A2C) For The CartPole Problem *Feb 2024 - May 2024*

Course project for Artificial neural networks/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 : Frédéric Kaplan

- Integrated ImageBind and Stable Diffusion model into a unified multimodal pipeline; built a benchmark dataset (instruments, food, clothing, paintings) with cross-modal inputs (text, image, audio) to evaluate the model's cultural comprehension and image-text coherence through generation and retrieval tasks.

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

Course project for Applied Data Analysis (Best 10 projects) | Instuctor: Bob 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*

VOLUNTEER

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

TECHNICAL STRENGTHS

Programming Languages: Python, JavaScript, R, SQL, C
Frameworks & Libraries: PyTorch, TensorFlow, pandas, scikit-learn, Spark, React
Tools: MATLAB, Git, LaTeX, SPSS
Languages: Chinese (Native); English (Proficient); French (Beginner)