Yiwei Liu

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

Education

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

National Institute of Informatics — Tokyo, Japan
Visiting Student, advised by Prof. Saku Sugawara — *Oct 2024 – Mar 2025 ; May 2025 – Now*

Beijing Institute of Technology (BIT) — Beijing, China
BSc in Statistics (GPA: 88/100, Major GPA: 90/100) — *Sept 2019 – June 2023*

Relevant Coursework:

Statistics: Probability Theory, Mathematical Statistics, Multivariate Statistical Analysis, Applied Stochastic Processes. **Mathematics:** Mathematical Analysis, Advanced Algebra, Optimization Methods, Function of Real Variables. **Computational Methods:** Machine Learning, Modern Natural Language Processing, Computational Social Media, Reinforcement Learning.

Research Interest

My long-term goal is to enable trustworthy human–AI collaboration in future agentic ecosystems. I'm particularly interested in sociotechnical challenges arising from emergent social behaviors – such as deception and collusion in multi-agent and human–AI interactions. I study both the behavioral patterns and underlying mechanisms of LLMs to understand how such challenges arise. I'm especially motivated by interdisciplinary methods that distill insights from cognitive science (e.g., Theory of Mind and decision-making modeling).

Publications And Technical Reports

- TactfulToM: Do LLMs Have the Theory of Mind Ability to Understand White Lies?
  Yiwei Liu, Emma Jane Pretty, Jiahao Huang, Saku Sugawara
  In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)
- A Visual Resolution for Syntactic Ambiguity
  Semester Project Report; conducted at IVIA lab, ETH Zürich.
- Business Model Construction Logic of Research-based Platforms *(Chapter 12)*
  Co-author, contributed to data analysis; conducted under Prof. Wei Wei at Peking University.
  China Machine Press, 2024 (ISBN: 978-7-111-73714-8)

Research Experience

Mechanistic Interpretability Perspective of Injecting Knowledge into LLMs — *Aug 2025 – Now*
Supervisor: Dr. Thibault Laugel | AXA GO Rev Research

- Investigated the robustness and specificity of knowledge editing methods (*ROME*, *MEMIT*, *GRACE*) for LLMs, identifying critical limitations and under-explored unintended behaviors such as revert-back effects resembling jailbreaking and emergent hallucination patterns.
- Utilized interpretability tools (*LogitLens*, *Patchscope*) to probe the activation spcace, tracing how knowledge edits propagate and degrade across model layers, informing the design of more robust intervention methods for safer alignment.

Theory of Mind Reasoning Evaluation of LLMs in White-Lie Understanding — *Oct 2024 – Mar 2025*
Supervisor: Prof. Saku Sugawara

- Designed *TactfulToM*, an English ToM benchmark (100 samples, 6.7k questions) designed to evaluate LLMs' understanding of white lies through complex social scenarios via systematic templates simulating prosocial deception.
- Revealed that even state-of-the-art LLMs underperform humans in white lies understanding and reasoning, particularly in understanding the emotional motivation behind it, raising a critical alignment question: should LLMs understand white lies merely to interpret human behavior, or also to potentially generate them?

Handling Uncertainty in LLMs: From Ambiguity Resolution to Utility-Driven Decision-Making

Evaluating LLM Decision-Making under Uncertainty via Utility Modeling *May 2025 – Now*
Supervisor: Prof. Saku Sugawara

- Developing a computational framework based on inverse preference inference to evaluate LLM decision-making under uncertainty, adapting Bayesian cognitive models to analyze model behavior in high-stakes domains (e.g., medical dialogue, psychotherapy).
- Integrating pragmatic reasoning models such as Rational Speech Acts (RSA) to guide utility-driven action selection, enabling LLMs to generate optimal clarification questions when complete contextual information is unavailable.

Syntactic Ambiguity: Detection and Visual Resolution *June 2024 – Dec 2024*
Supervisors: Prof. Frédéric Kaplan, Prof. Menna El-Assady, Robin Chan

- Built a k-best constituency parser and interactive interface to detect and visualize syntactic ambiguity through alternative interpretation comparison, motivated by initial evaluations showing that LLMs struggle to reliably identify ambiguous sentences.

Internship Experience

Beijing Xianfeng Changqing Venture Capital (K2VC) Beijing, China
Investment Research Intern | Supervisor: Runxin Yang *Dec 2022 – June 2023*

- Conducted evaluations of LLM startups, tracking paradigm shifts and identifying research gaps in LLM development.

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

- Engaged with leading AI labs in China to track frontier developments in deep learning; performed technical evaluation on investment targets including TensorOpera AI and Thewake Systems, assessing research-to-product pathways.

JD.com Explore Academy (AI Lab) Beijing, China
Algorithm Engineer Intern| Supervisor: Dr. Chang Li *June 2021 – Aug 2021*

- Developed a Proof-of-Concept for JD.com × Nvidia Omniverse Virtual Retailer that collects eye-tracking and interaction data and applies multimodal deep learning methods for behavior prediction to optimize shelf layout; contributed to internal AI research strategy, focusing on multimodal and trustworthy AI for e-commerce.

Selected Course Projects

Developed a Specialized LLM Tutor for EPFL Course *Feb 2024 – May 2024*
Modern Natural Language Processing (CS-552) | Instructor: Antoine Bosselut

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

Extending Text2Image Model to Multimodal Model *Sept 2023 – Dec 2023*
Foundation of Digital Humanities (DH-405) | Instructor: Frédéric Kaplan

- Integrated ImageBind and Stable Diffusion into a multimodal pipeline; developed a benchmark with cross-modal inputs across major cultural groups to evaluate cultural understanding and image–text coherence, revealing limited generalization in cultural concept understanding.

Honors and Awards

- Awarded "Gintong Talent Fellowship", Renmin University of China — 35 students selected annually *2022*
- Awarded Academic Excellence 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, Linux.
- **Languages:** English (Proficient), Chinese (Native), French (Beginner), German (Beginner)