

Weilun Xu

✉ weilun.xu@epfl.ch

🐙 [GitHub](#)

🎓 [Google Scholar](#)

🌐 [atopos.space](#)

EDUCATION

École Polytechnique Fédérale de Lausanne

09/2024 – 08/2027 (expected)

M.Sc. Digital Humanities

GPA: $\approx 5.5 / 6$

University College London

09/2023 – 12/2023

M.Sc. Machine Learning

Left to join EPFL

University of Melbourne

03/2019 – 07/2023

Bachelor of Science, Mathematics and Statistics

WAM: 87.2 (First Class Honours)

Diploma in Computing (WAM: 90.8); Diploma in Music (WAM: 82.2)

Relevant coursework: Bayesian Deep Learning, Probabilistic & Unsupervised Learning, Natural Language Processing, Machine Learning, Modern Applied Statistics, Stochastic Modelling, Real Analysis, Cognitive Psychology.

Awards: Dean's Honour List (Top 3%), International Undergraduate Scholarship (Top 1%).

PUBLICATIONS

Mechanistic Interpretability

Xu, W., & Sun, Z. (2026). Steering Vectors for Competing Ethical Frameworks Are Interchangeable: A Competing-Target Test for Activation Steering. *Submitted to ACL Rolling Review, October 2026*.
arXiv:2610.19359.

Xu, W., Rusnak, A., & Kaplan, F. (2026). Ethical Frameworks in LLMs Route Through a Shared Endorsement Subspace. *Mechanistic Interpretability Workshop at ICML 2026*.

Xu, W. (2026). Width-to-Depth Ratio Tracks Readout Misalignment in Language Models. *Mechanistic Interpretability Workshop at ICML 2026*.

Cognitive Science of Humans and Machines

Kemp, C., **Xu, W.,** & Cropper, S. J. (2026). A Bayesian account of the phantom lighthouse illusion. *Open Mind: Discoveries in Cognitive Science*, 1–16.

Xu, W., Cropper, S. J., & Kemp, C. (2023). A Bayesian account of two visual illusions involving lighthouse beams. *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Rashidi, S., **Xu, W.,** Lin, D., Turpin, A., Kulik, L., & Ehinger, K. (2023). An active foveated gaze prediction algorithm based on a Bayesian ideal observer. *Pattern Recognition*, 143, 109694.

Computational Musicology

Xu, W., Hall, E., & Rohrmeier, M. (2026). Diachronic Modeling of Tonal Coherence on the Tonnetz Across Classical and Popular Repertoires. *Proceedings of the 23rd Sound and Music Computing Conference*.

Xu, W., Dai, T., Goudet, O., & Wang, X. (2026). Acoustic Overspecification in Electronic Dance Music Taxonomy. *Proceedings of the 23rd Sound and Music Computing Conference*.

RESEARCH EXPERIENCE

Mechanistic Interpretability

Representing and Steering Ethical Frameworks in LLMs

05/2025 – present

Advisors: Alexander Rusnak & Prof. Frédéric Kaplan, EPFL

With Zipeng Sun (McGill / Mila)

- Found, across 11 open-weight LLMs and six moral systems, that framework directions are distinguishable yet steer through one shared *endorsement subspace*, where one direction often substitutes for another.
- Showed that steering shifts the invoked framework more than the chosen action, and that SFT, not preference learning, does most of the reshaping of this geometry.
- Introduced a *competing-target test*: Kantian and utilitarian steering vectors proved interchangeable, because a shared request to apply some framework dominated both, whereas prompting kept them apart; a vector's gain on its named target does not show what it controls.

Readout Misalignment in Language Models

01/2026 – present

Independent research, EPFL

- Introduced PGA, a null-calibrated score testing whether a layer’s geometry lies in the output map’s directions, as logit-lens-style interpretability assumes.
- Found a late-layer readout misalignment in narrow models, invisible to the training loss, whose severity is ordered by width-to-depth ratio d/L rather than parameter count across seven Pythia models.
- Showed that the misalignment masks rather than erases predictive structure, which removing a few dominant directions recovers: a failed reading at such layers reflects the readout, not missing knowledge.

LLMs for Social Science

Human Attitude Estimation with LLMs

07/2026 – present

AXA Group Operations, Lausanne

Research Intern

Mentors: Rubén Garzón, Vincent Grari, Marcin Detyniecki & Michael Bernstein

- Found useful LLM signal in aggregate attitudes alongside weak prediction of individual responses, motivating a shift from silicon samples to silicon aggregates.
- Extended a Fay–Herriot estimator with LLM forecasts in the mean and similarities between survey questions in the covariance, calibrated against human responses.
- Lowered mean squared error beyond forecasts alone in GSS development experiments; raised minimum interval coverage from 78% to 89.5% across 24 simulation settings.

Cognitive Science of Humans and Machines

A Bayesian Account of Two Visual Illusions

08/2022 – 04/2023

Advisors: Prof. Charles Kemp & Prof. Simon Cropper, University of Melbourne

- Built a Bayesian model explaining bent lighthouse beams and the “phantom lighthouse” illusion from natural scene statistics, predicting that the phantom appears only beyond about 75 m from the true source.
- Showed in field and virtual-reality experiments with human observers that the illusion emerges as predicted.

Active Foveated Gaze Prediction with a Bayesian Ideal Observer

06/2021 – 06/2022

Advisor: Prof. Kris Ehinger, University of Melbourne

- Adapted a Bayesian ideal-observer model for foveated fixation prediction that simulates peripheral vision loss with eccentricity-based max pooling and chooses each fixation to maximise information gain.
- Without eye-movement training, performed comparably to fixation-trained models on two benchmarks.

Visual Cue Insufficiencies and Emotion Recognition

02/2022 – 06/2022

Advisors: Dr. Tilman Dingler & Dr. Zhanna Sarsenbayeva, University of Melbourne

- Built a real-time emotion-detection video conferencing interface and ran a 40-participant facial-occlusion experiment with physiological sensing.

Computational Musicology

Diachronic Modeling of Tonal Coherence on the Tonnetz

09/2025 – 03/2026

Advisors: Dr. Edward Hall & Prof. Martin Rohrmeier, DCML, EPFL

- Proposed tonal focus and tonal connection as partially independent dimensions of tonal coherence; across 2,800+ classical and popular pieces, popular music is higher in focus and classical in connection.

Acoustic Overspecification in Electronic Dance Music Taxonomy

04/2025 – 02/2026

Advisor: Xiaoxuan Wang, DCML, EPFL

- Designed an interpretable acoustic feature space for EDM and showed that its 35-genre commercial taxonomy is acoustically overspecified: unsupervised clustering finds 17–20 natural acoustic families.

Information Retrieval

Reliability of Listwise Learning-to-Rank

06/2022 – 11/2022

Advisors: Dr. Lida Rashidi & Prof. Justin Zobel, University of Melbourne

Undergraduate Thesis

- Tested whether gains in MAP and NDCG@10 track modelled user utility (RBP), comparing learning curves for four rankers across three LETOR benchmarks using RankLib.

TECHNICAL SKILLS

Languages: Python, R, JavaScript, $\text{\LaTeX}$, SQL, C, Java

ML/AI: PyTorch, Hugging Face Transformers, scikit-learn, NumPy, SciPy, pandas, Matplotlib, librosa

Probabilistic / Bayesian: PyMC, custom MCMC and variational inference, Gaussian processes

Tools: Git, Linux, Docker, React, Node.js

TEACHING

Teaching Assistant, University of Melbourne

2021 – 2023

- *Machine Learning; Elements of Data Processing; Design of Algorithms.*