Xueyan Li

lixuey@ethz.ch Personal Website LinkedIn GitHub

Education

ETH Zurich - PhD in Machine Learning 2024-2028

- Funded by the CLS program, joint collaboration between the Max Planck Institute and ETH Zurich.
- Primary supervisor: Dr Jonas Geiping. Secondary supervisor: Prof Mrinmaya Sachan.

University of Cambridge – MPhil in Machine Learning and Machine Intelligence 2022-2023

- Overall mark: 76.6/100 & Dissertation mark: 87.0/100 (Transcript)
- Relevant courses: Computer Vision, Spoken Language Generation and Recognition, Machine Translation and Dialogue Systems, Probabilistic ML, Deep Learning and Structured Data

Imperial College London – MEng in Bioengineering 2018-2022

- First Class Honors 74.4/100 (Transcript)
- Relevant courses: Signals and Control, Image Processing, Hearing and Speech Processing, Digital Biosignal Processing, Probability and Statistics, Reinforcement Learning

Raffles Institution - Singapore 2016 - 2017

- Recipient of Singapore Ministry of Education SM1 scholarship for 4 years
- Singapore-Cambridge GCE A Level: AAAA; SAT score: 1500 (99th percentile)

Research Interests

- LLM uncertainty
- Sampling strategies
- Diversity vs accuracy trade-off
- Chain-of-thought reasoning
- Token-level calibration
- Multi-step generation and backtracking
- Interpretable and trustworthy systems
- Multi-modal integration

Research Papers

Xueyan Li, Guinean Su, Mrinmaya Sachan and Jonas Geiping. **Sample Smart, Not Hard: Correctness-First Decoding for Better Reasoning in LLMs**. *Under review at ICLR 2026*. Preprint:dddd Code.

- We challenge the common assumption that high model uncertainty warrants more exploration in order to generate different reasoning paths that would eventually lead to a correct final answer.
- We show that epistemic and aleatoric uncertainty are confounded in literature, and propose a counter-intuitive method to restrict sampling when model uncertainty is high. We test on long-form reasoning tasks with significant improvement when compared to similar samplers in literature.
- We propose a novel but straightforward method to assess step-wise calibration. We make a surprising finding that at highly uncertainty steps, larger models are not more correct than smaller models.

Xueyan Li and Jonas Geiping. **Confidence-As-Context: Persistent Uncertainty Signals for Autoregressive Generation.** 2025. Preprint. Code.

- Step-wise token uncertainties contain rich information on model confidence but are discarded after each step. We design a novel training and inference strategy and make these uncertainty signals persist throughout generation. All future generations are dependent on all past generation uncertainties.
- This method should allow models to self-assess accuracy and back-track, pause or continue, therefore achieving better chain-of-thought performance. Work in progress.

Xueyan Li and Bill Byrne. **Incorporating Vision Encoders into Retrieval Augmented Visual Question Answering.** 2023. Master's thesis. Short paper. Code.

- MPhil thesis at University of Cambridge. Obtained mark 87.0/100.
- We find limitations in existing retrieval augmented visual question answering methods, and propose a novel pipeline that make use of common methods in literature.
- Added various visual encoders (CLIP, BLIP2, InstructBLIP) to retrieval-augmented large language models. Performance on the OKVQA dataset outperforms all other retrieval-based models.
- Frozen LLMs (e.g. GPT-3.5) are used to choose a final answer from document-based answer candidates with in-context learning. Performance outperforms all other in-context methods.

Key Projects

Adaptive Head for Step-Wise Temperature or Correctness Signals March - June 2025

- We train a transformer or MLP layer to output step-wise temperature or correctness signals.
- We deep dive into an existing method ATS and find multiple implementation errors. We fix these errors and improve them in terms of generalizability and versatility.
- Details can be found in the project page. Work in progress.

Gait segmentation using low dimension output embedding mapping in CNN Oct 2021 – June 2022

- MEng thesis supervised by Dr Lance Rane. (Paper)
- Proposed a novel approach based on output embedding mapping to encode the cyclical nature of a gait cycle.
- Used knowledge distillation, weight pruning and quantization to compress CNN to have minimal latency and optimal performance on wearable hardware with limited memory.
- Focused on the generalizability of gait classification on pathological gait data collected in house.

Performance of various loss functions for Siamese neural networks Jan 2023

- Compared binary cross-entropy, contrastive and triplet loss functions to classify 20 classes of images from Tiny ImageNet. Compared ViT vs CNN architecture in terms of training data required, preservation of features across layers, and the effect of skip connections. (Paper)

Interpretability of polysemantic neurons in autoencoders March 2023

- Investigated the phenomenon of superposition as described in the paper "Toy Models of Superposition."
- Understand how autoencoders can disentangle high dimensionality features with limited number of neurons. (Poster)

Grammars of Action July – Sept 2020

- Summer research project at the Imperial College London Brain and Behaviour Lab.
- Annotated motor behaviours of subjects performing daily activities to train a Multinomial Hidden Markov Model using the Expectation-Maximization algorithm.
- Explored usage of modular reinforcement learning and sequitur algorithm to model multitask behaviours.

Hospital database web app development with Java Nov – Dec 2020

- Collaborated with the NHS to develop an app for doctors to manage information when changing shifts.
- Configured Spring Boot to use PostgreSQL database and built Restful CRUD API.
- Continuously deployed React app to Netlify.
- Familiarized with object-oriented programming paradigms and Agile concepts. (code)

Work Experience

Deep Learning intern at Oxehealth - Oxford, UK July - Sept 2022

- Worked with deep residual networks that evaluate patient health from physiological data.
- Used transfer learning to train model on public data to augment and improve performance on in house data.
- Improved data pre-processing pipeline and tested various hyperparameters, loss functions and model architectures. Built training pipeline and data readers.

Software Engineering intern at Deutsche Bank - London, UK June - Aug 2021

- Familiarized with production pipeline and technology stack like Hadoop, Artifactory, Fabric etc.
- Researched the impact of Covid-19 on different industries and predicted stock performance using CNNs.
- Created video popup function in Typescript for an internal web app.
- Used React to create a mock-up of a bug tracking web app, and deployed to Google App Engine.

Technology intern at Goldman Sachs - London, UK March 2020

- Researched and gave a presentation on the role of functional programming in finance.
- Familiarized with collaborative processes and project management with Jira, Git and Agile.
- Networked with employees to learn about different divisions and company culture.

Additional Information

Leadership positions: Imperial College Women in Business Society Marketing Manager, Imperial College Women in SET Industrial Liaison
Languages: Mandarin (native), English (fluent)
Programming languages: Python, C++, Java, Matlab, React
Interests: cooking, baking, bouldering, hiking, weekend trips