Lexin Zhou

lexinzhouds@gmail.com — https://lexzhou.github.io/

I spend my day thinking about (i) creating robust evaluation methods that provide robust descriptions of AI's capabilities and risks, and (ii) finding ways to positively shape the predictability and reliability of AI; I have a special interest in foundation models.

Education

University of Cambridge — *Cambridge, UK*
MPhil of Advanced Computer Science (NLP Track) — *2023 – 2024*
- Grade: Distinction (GPA = 4.0/4.0)
- Funded by Open Philanthropy Long-Term Future Scholarship

Universidad Politécnica de Valencia (UPV) — *Valencia, Spain*
Bachelor of Science, Data Science — *2019 – 2023*
- Grade: 9.54/10 (ranked the 1st in the cohort)
- Funded by the Spanish Ministry of Education

Representative Publications

[1] **Lexin Zhou**, Fernando Martínez-Plumed, Wout Schellaert, Yael Moros-Daval, Cèsar Ferri, José Hernández-Orallo. "Scaled-up, Shaped-up, but Letting Down? Reliability Fluctuations of Large Language Model Families", 2024. *Nature*, to appear.

[2] **Lexin Zhou**, Youmna Farag, Andreas Vlachos. "An LLM Feature-based Framework for Dialogue Constructiveness Assessment", 2024. Under review at *EMNLP 2024* (status: avg. review score=4.17/5; top 1% of the submissions).

[3] **Lexin Zhou**, Pablo Moreno-Casares, Fernando Martínez-Plumed, […], José Hernández-Orallo. "Predictable Artificial Intelligence", 2023. Under review at *Artificial Intelligence Journal* (status: received a major revision).

Other Publications

[4] **Lexin Zhou**, Fernando Martínez-Plumed, José Hernández-Orallo, Cèsar Ferri, Wout Schellaert. "Reject Before You Run: Small Assessors Anticipate Big Language Models", 2022. *Evaluation Beyond Metrics Workshop@IJCAI-2022*.

[5] Anthony G Cohn, José Hernández-Orallo, Julius Sechang Mboli, Yael Moros-Daval, Zhiliang Xiang, **Lexin Zhou**. "A Framework for Categorising AI Evaluation Instruments", 2022. *Evaluation Beyond Metrics Workshop@IJCAI-2022*.

[6] Nekane Romero-Garcia, **Lexin Zhou**, […], Carlos Sáez. "Machine Learning Uncovers Blood Test Patterns Subphenotypes at Hospital Admission Discerning Increased 30-day ICU Mortality Rates in COVID-19 Elderly Patients", 2022. *The 42nd International Symposium on Intensive Care & Emergency Medicine*.

[7] **Lexin Zhou**, Nekane Romero, Juan Martínez-Miranda, J Alberto Conejero, Juan M García-Gómez, Carlos Sáez. "Subphenotyping of Mexican Patients With COVID-19 at Preadmission to Anticipate Severity Stratification: Age-Sex Unbiased Meta-Clustering Technique", 2022. *JMIR Public Health and Surveillance*.

Research Experience

University of Cambridge
MASTER'S DISSERTATION (advised by Prof. Andreas Vlachos and Dr. Youmna Farag) — *Jan 2024 – Jun 2024*
- Developed a novel framework that leverages LLMs to generate rich linguistic features to train human-interpretable feature-based models that achieve state-of-the-art performance with robust prediction mechanisms across several dialogue constructiveness prediction tasks [2].

Meta AI
INDEPENDENT CONTRACTOR — *Jan 2024 – Mar 2024*
- Adversarial testing Meta's new foundation models in the red team, to elicit model vulnerabilities that may lead to unsafe behaviours.

Kruger AI Safety Lab
RESEARCH INTERN (advised by Dr. Gabriel Recchia and Prof. Jose Hernandez-Orallo) — *June 2023 – Sep 2023*
- Experimented the utility of *assessor* models for scalable oversight and alignment, anticipating unsafe behaviour of LLMs and providing feedback to make LLMs more aligned, taking features derived from <task instance, subject system>.

VRAIN, UPV
RESEARCH ASSISTANT (advised by Prof. Jose Hernandez-Orallo) — *Jan 2022 – May 2023*

- Evaluated the reliability evolution of large language model families (e.g., GPT, LLaMA, BLOOM). The result is presented in a paper accepted at *Nature* [1].
- AI predictability and evaluation research on failure prediction of LLMs (e.g., GPT-3, BIG-G) in downstream tasks at instance-level [4].
- Helped designed and improved a novel rubric for categorising AI Evaluation Instruments [5].

OpenAI

INDEPENDENT CONTRACTOR *Sep 2022 – Mar 2023*

- Evaluated OpenAI's new foundation models in the red team, including adversarial testing on GPT-4, red teaming, evaluating capabilities of GPT-4, optimizing GPT-4 with human feedback, and writing reports. Part of the work is presented in [1].

Joint Research Centre, European Commission

CONTRACTED EXPERT *Jul 2022 – Aug 2022*

- Integrated instance-level data from recent AI benchmarks into the AICollaboratory of AI Watch.
- Evaluated the predictive power of GPT-4-prompted meta-features (as proxies of task difficulty) of task instances, to evaluate the extent to which performance of LLMs can be anticipated and explained through meta-features.

Biomedical Data Science Lab, UPV

RESEARCH COLLABORATOR *Jun 2020 –December 2021*

- Designed a novel bias mitigation clustering methodology that provides less biased results across subgroups [7].
- Worked with Valencian Clinic Hospital and University Hospital 12 de Octubre de Madrid i + 12 to analyse Electronic Health Records of patients with unsupervised learning, facilitating an early risk stratification and help decision-making in the resource allocation process [6].

Other Experience

Leverhulme Centre for the Future of Intelligence & Centre for the Study of Existential Risk

STUDENT FELLOW (advised by Dr. Seán Ó hÉigeartaigh) *Oct 2023 - Present*

- Research student fellow working on AI evaluation and scalable oversight at the CFI's AI: Futures and Responsibility research program.

Co-organising the 1st workshop on **Predictable AI**

ORGANISING COMMITTE *March 2023*

- Co-organising the 1st workshop of "*Predictable AI: Evaluation, Anticipation and Control*", an initiative funded by the Future of Life Institute, with invited speakers from Deep Mind, Microsoft Research, Mila, Cambridge and EC JRC. The topics ranged from scaling laws, control, liability and future risks to cognitive and robust evaluation, assessors, co-operative conditions, uncertainty estimation. In [3], we summarise the key ideas of what Predictable AI is, the main questions, hypotheses and challenges, as well as identifying paths towards AI predictability and the potential impact of this emergent field.

Honors & Awards

2023 **Open Philanthropy Long-Term Future Scholarship** supporting my one-year master's studies, Open Philanthropy
2023 **1st Prize for the research publication with the highest impact factor in 2022**, ITACA Institute, UPV
19'-23' **"Best Academic Record" Awards** (ranked 1st in the undergraduate cohort of the data science degree), UPV
2022 **Undergraduate Research Collaboration Fellowship**, Ministry of Education, Government of Spain
2022 **"Artificial Intelligence and the Future of Skills" Research Fellowship**, VRAIN & OECD
2022 **Santander Bank Studies Progress Scholarship** (top 0.0001% in university), Santander Bank

Skills

Language: English (TOEFL): 109 (R28, L29, S26, W26), Spanish & Chinese: Mother Tongue, Catalan: Limited Working Proficiency.
Research Interests: AI Evaluation, Predictable AI, AI Reliability, AI Safety, Benchmark Validity, General-Purpose AI, Language Models.
Other Interests: Outside of AI evaluation I spend my time playing piano, hiking with friends, traveling with loved ones, or reading about miscellaneous science/philosophical content.

Miscellaneous

- Reviewer at: AMMAS 2023, ACL 2023
- Interviewed for Final Times article "OpenAI's red team: the experts hired to 'break' ChatGPT"
- Invited talk on "An LLM Feature-based Framework for Dialogue Constructiveness Assessment" at Toshiba Cambridge