Lexin Zhou

✉ Email: lexinzhouds@gmail.com — 🌐 Homepage: https://lexzhou.github.io/

I mostly spend my day thinking about (i) designing robust evaluation methods that offer explanatory and predictive power of AI's capabilities and risks, and (ii) finding pathways to positively shape their reliability and predictability. I'm also interested in psychometrics, AI's social implications, safety, and cognitive sciences; I am especially intrigued by LLMs.

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
- Advisor: Andreas Vlachos.

Universidad Politécnica de Valencia   *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
- Advisor: Jose Hernandez-Orallo

Representative Publications

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

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

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

Other Publications

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

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

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

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

Research & Industry Experience

Meta AI
Independent Contractor   *Jan 2024 – Mar 2024*

Kruger AI Safety Lab
Research Intern (advised by Dr. Gabriel Recchia & Prof. Jose Hernandez-Orallo)   *June 2023 – Sep 2023*

VRAIN, Universidad Politécnica de Valencia

Undergraduate Research Assistant (advised by Prof. Jose Hernandez-Orallo)   *Jan 2022 – May 2023*

OpenAI

Independent Contractor   *Sep 2022 – Mar 2023*

Joint Research Centre, European Commission

Expert Contract   *Jul 2022 – Aug 2022*

Biomedical Data Science Lab, Universidad Politécnica de Valencia

Undergraduate Research Collaborator (advised by Dr. Carlos Sáez)   *Jun 2020 –December 2021*

Honors & Awards

2023   Open Philanthropy Long-Term Future Scholarship, Open Philanthropy

19'-23' "Best Academic Record" Awards (ranked the 1st in the BSc data science cohort), Universidad Politécnica de Valencia

2023   1st Prize for the research publication with the highest impact factor in 2022, ITACA Institute, 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 0.001% in university), Santander Bank

Miscellaneous

- Language: English, Spanish, Chinese, and Catalan.

- Leisure Activities: Outside the realm of science, I spend my time playing piano, swimming, hiking with friends, traveling with loved ones, or enjoy reading about miscellaneous science/philosophical content.

- Conference Reviewer: AMMAS 2023, ACL 2023.

- Interview: OpenAI's red team: the experts hired to 'break' ChatGPT, *Final Times*.

- Invited Talk: "An LLM Feature-based Framework for Dialogue Constructiveness Assessment" at Toshiba Cambridge.

- Conference Committee: The 1st workshop on Predictable AI.