ALINA LEIDINGER

✉ a.j.leidinger@uva.nl aleidinger.github.io Linkedin Google Scholar @alinaleidinger

PERSONAL PROFILE

I am a final year PhD student at the University of Amsterdam where I research robustness, bias and values in large Language Models. In my published works, I take an interdisciplinary approach drawing on computer science, ethics, logic, and media studies to tackle issues such as LLM instability and stereotyping.

EDUCATION

ILLC, University of Amsterdam, NL *09/21 - 09/25*
PhD candidate in NLP
- Topic: Implicit bias and stereotypes in Large Language Models
  Advisors: Dr. Katia Shutova & Prof. Dr. Robert van Rooij

University of Augsburg, DE *07/20 - 07/21*
Research Assistant in HCI, NLP

Technical University of Munich, DE *10/16 - 08/19*
MSc Mathematics in Data Science *High Distinction*
- Thesis topic: Mathematical Analysis of Neural Networks, 1.0 (top grade)

Imperial College London, UK *10/13 - 07/16*
BSc Mathematics *First Class Honours*
- Thesis topic: Robust Joint and Individual Variance Explained, CVPR 2017

WORK EXPERIENCE

HuggingFace *09/23 - 05/24*
Research collaboration on CIVICS cultural values dataset

LMU Munich, CIS *09/23 - 10/23*
Research visit: bias from pretraining to alignment
Host: Prof. Dr. Hinrich Schütze

BMW *06/18 - 12/18*
Machine Learning Intern

TUM, Chair of Computer Vision *07/17 - 09/17*
Student research assistant

MunichRe, MEAG *07/15 - 08/15*
Statistical Modelling Intern

TECHNICAL SKILLS

- **Programming** Python (PyTorch, transformers, numpy, pandas, sklearn, matplotlib, seaborn), R
- **Tools** git, slurm, LaTeX

AWARDS & PRIZES

- **Finalist at UvA 3MT competition** *08/22*
  Science communication competition for PhD students
- **ScienceHack at TUM** *12/19*
  First prize at hackathon for building a CO_2 Counter
- **Deutschlandstipendium** *'17 - '18*
  Scholarship awarded by German Federal State to $< 2\%$ of students
- **PreDoc programme at TUM** *'16 - '18*
  PhD preparation programme for outstanding mathematics students

PUBLICATIONS

M. Thaler, A. Köksal, **A. Leidinger**, A. Korhonen, H. Schütze. 2024. Bias Propagation in LLMs: Tracing Gender Bias from Pre-training Data to Alignment. *Under review.*

G. Pistilli*, **A. Leidinger***, Y. Jernite, A. Kasirzadeh, A. Luccioni, M. Mitchell. 2024. CIVICS: Building a Dataset for Examining Culturally-Informed Values in Large Language Models. *ACM/AAAI Artificial Intelligence, Ethics, and Society*

A. Leidinger and R. Rogers. 2024. How are LLMs mitigating stereotyping harms? Learning from search engine studies. *ACM/AAAI Artificial Intelligence, Ethics, and Society*

I. Solaiman*, Z. Talat*, W. Agnew, L. Ahmad, D. Baker, S.L. Blodgett, C. Chen, H. Daumé III, J. Dodge, I. Duan, E. Evans, F. Friedrich, A. Ghosh, U. Gohar, S. Hooker, Y. Jernite, R. Kalluri, A. Lusoli, **A. Leidinger**, M. Lin, X. Lin, S. Luccioni, J. Mickel, M. Mitchell, J. Newman, A. Ovalle, M.T. Png, S. Singh, A. Strait, L. Struppek, A. Subramonian. 2023. Evaluating the social impact of generative AI systems in systems and society. *Forthcoming in Hacker, Engel, Hammer, Mittelstadt (eds), Oxford Handbook on the Foundations and Regulation of Generative AI. Oxford University Press.*

A. Leidinger, R. van Rooij, E. Shutova. 2024. Are LLMs classical or nonmonotonic reasoners? Lessons from generics. *ACL 2024 (main)*

A. Leidinger, R. van Rooij, E. Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? *Findings of EMNLP 2023*

G. Starace, K. Papakostas, R. Choenni, A. Panagiotopoulos, M. Rosati, **A. Leidinger**, E. Shutova. 2023. Probing LLMs for Joint Encoding of Linguistic Categories. *Findings of EMNLP 2023*

A. Leidinger and R. Rogers. 2023. Which Stereotypes Are Moderated and Under-Moderated in Search Engine Autocompletion? *ACM FAccT 2023*

O. van der Wal, D. Bachmann, **A. Leidinger**, L. van Maanden, J. Zuidema, K. Schulz. 2022. Undesirable biases in NLP: Averting a crisis of measurement. *JAIR*

C. Sagonas, I. Panagakis, **A. Leidinger**, S. Zafeiriou. 2017. Robust Joint and Individual Variance Explained. *CVPR 2017*

TALKS & PRESENTATIONS

- Workshop GenBench, collocated with EMNLP *12/23*
- Workshop on Ethical AI, Comète (Inria Polytechnique) *11/23*
- Workshop on generative AI and search engines, HAW Hamburg *09/23*
- Interview at Tech Policy Press podcast *08/23*
- Invited talk at Civic AI Lab, UvA *02/23*
- Flash talk for deans of the faculties of humanities and science, UvA *02/23*
- Presentation at ELLIS PhD Symposium, University of Alicante *09/22*

TEACHING EXPERIENCE

- Teaching assistant: Advanced Topics in Computational Semantics *Spring '22, '23, '24*
- Teaching assistant: Natural Language Processing 1 *Fall '22, '23*
- Teaching assistant: Basic Probability: Programming *Fall '21*
- Master thesis supervisor: Debiasing LMs using Influence Functions. *10/22 - 06/23*

SERVICE

- Reviewing: *ACL, EMNLP, ARR, COLING
- Co-organiser of Computational Linguistics Seminar at ILLC *04/22 - ongoing*

LANGUAGES

German: C2 | English: C2 | French: B2 | Italian: B2 | Dutch: B1