KRISTIN WITTE

Doctoral Researcher in Behavioral Evaluation of LLMs & Computational Modeling

@ kristin.witte@helmholtz-munich.de   kristinwitte.github.io   kristin-witte-a356b7161   KristinWitte

Researcher bridging computational neuroscience and AI safety, with 8+ years of experience designing and evaluating experiments. Skilled in Python, Bayesian modeling, and LLM fine-tuning, with interests in deceptive alignment, behavioral evaluations of LLMs, and control of frontier models. Focused on building robust and trustworthy AI systems.

KEY EXPERIENCE

Doctoral Researcher

Helmholtz Munich · Ludwig-Maximilians-University

📅 Oct. 2022 – Ongoing   📍 Munich, Germany

- Led a 3-year research program on robustness and interpretability in human decision-making models (Scientific Reports, 2025)
- Built Bayesian and Gaussian Process models to characterize uncertainty in learning processes (OSF preprint)
- Designed behavioral evaluation protocols for LLMs to assess robustness, consistency, and contextual cues shift output behavior, relevant to detecting deceptive alignment patterns (npj Digital Medicine, 2025, arXiv, 2024)
- Fine-tuned LLMs (PEFT) and analyzed emergent failure modes
- Released reproducible evaluation pipelines and collaborated with interdisciplinary teams

Graduate Researcher

University College London · Max Planck Institute for Biological Cybernetics

📅 Oct. 2020 – Sep. 2022   📍 London, UK; Tuebingen, Germany

- Designed and implemented large-scale online experiments (JS, HTML) to test causal links between worry and exploratory decision-making
- Applied hierarchical Bayesian and Gaussian Process modeling to analyze behavioral data and evaluate exploration strategies under uncertainty

Summer Intern

Massachusetts Institute of Technology

📅 Jul. 2019 – Aug. 2019   📍 Cambridge, MA

- Delivered quantitative insights into affective influences on information processing (Journal of Neuroscience, 2021)

KEY PUBLICATIONS

For a complete list of publications, see Google Scholar

Journal Articles

- Z. Ben-Zion, **K. Witte**, A. K. Jagadish, *et al.*, “Assessing and alleviating state anxiety in large language models,” *npj Digital Medicine*, vol. 8, no. 1, pp. 1–6, 2025.
- M. Binz, ..., **K. Witte**, ..., and E. Schulz, “A foundation model to predict and capture human cognition,” *Nature*, pp. 1–8, 2025.
- **K. Witte**, M. Thalmann, and E. Schulz, “Model-based exploration is measurable across tasks but not linked to personality and psychiatric assessments,” *Scientific Reports*, 2025.

SKILLS

Python · R · JAX · MATLAB · Git · JavaScript · HTML · LaTeX

LLM Safety Evaluation · Hierarchical Bayesian Modelling · LLM Fine-Tuning · RL · Uncertainty Quantification · Privacy-Preserving ML · Model Interpretability

Experiment Design & Analysis · Research Leadership & Communication · Statistical Analysis

EDUCATION

Ph.D. Psychology

Ludwig-Maximilians-University

📅 2022 – present   📍 Munich, Germany

MSc. Neural and Behavioural Science

University of Tuebingen

📅 2020 – 2022   📍 Tuebingen, Germany

B.Sc. Psychology

Radboud University

📅 2016 – 2019   📍 Nijmegen, Netherlands

LANGUAGES

English (C2), German (native), French (B2)

REFEREES

Dr. Eric Schulz
@ Helmholtz Munich
✉ eric.schulz@helmholtz-munich.de
PhD Supervisor

Prof. Dr. Quentin Huys
@ University College London
✉ q.huys@ucl.ac.uk
Master Thesis Supervisor