Phongsakon Mark Konrad

+45 55 24 01 58 phongsakon@outlook.dk phomarkon.github.io
linkedin.com/in/phongsakonmarkkonrad github.com/phomarkon Google Scholar
Sønderborg, Denmark

Profile

Independent researcher in mechanistic interpretability of large language models. I trace the internal circuits behind sandbagging, demographic dishonesty, and moral-framing effects in open-weight models, and I ablate them. Adjacent threads cover calibrated uncertainty, the theoretical limits of self-prediction, the elicitation gap in automated red-teaming, and a benchmark proposal for evaluating interpretability methods themselves. Most of this work is solo, on a single Apple Silicon MacBook, against models small enough to instrument completely. Final-year BSc Software Engineering student at the University of Southern Denmark and incoming MPhil student in Machine Learning and Machine Intelligence at the University of Cambridge (October 2026), with an exchange semester at HKUST. Background includes four years in the German Armed Forces and three founder roles in fitness AI, ed-tech, and neurodivergent-friendly consumer software. Born in Thailand, raised in Germany, currently based in Denmark.

Education

University of Cambridge Oct 2026 (Incoming)
MPhil in Machine Learning and Machine Intelligence Cambridge, United Kingdom

The Hong Kong University of Science and Technology (HKUST) Sep 2025 – Dec 2025
Exchange Semester, Dept. of Computer Science and Engineering Hong Kong SAR, China

- Specialisation courses: COMP4211 Machine Learning, COMP4471 Deep Learning in Computer Vision, COMP4901B Large Language Models, COMP4901Z Reinforcement Learning.
- Postgraduate course: COMP6411D Data Visualisation.

University of Southern Denmark (SDU) Sep 2023 – Jun 2026
Bachelor of Science (BSc), Software Engineering Sønderborg, Denmark

- Expected graduation: June 2026. Current GPA: $\approx 10.6/12$ ($\approx 3.7/4.0$ US GPA).
- Practical, project-centric curriculum with mandatory semester-long team projects on complex, data-intensive software systems in domains such as IoT and AI, often in collaboration with industry partners.

Publications

Mechanistic Interpretability and AI Safety

1. **P. M. Konrad** and S. Ayvaz (2026). *When does chain-of-thought improve safety? Evidence from 18 models across 5 families.* Under review, *Conference on Language Modeling (COLM) 2026.*
2. **P. M. Konrad** (2026). *How Language Models Learn to Deceive: Anatomy of a Sandbagging Circuit.* Working paper, 2026.
3. **P. M. Konrad** (2026). *Differential Dishonesty: Language Models Encode User Demographics and Deviate from Their Own Beliefs Accordingly.* Working paper, 2026.
4. **P. M. Konrad** (2026). *Shingeki no Features: Are Moral Framing Effects in LLMs Shallow or Deep?* Working paper, 2026.
5. **P. M. Konrad** (2026). *Every Mirror Has a Blind Spot: A Fixed-Point Theory of Irreducible Self-Ignorance.* Working paper, 2026.
6. **P. M. Konrad** (2026). *Stop Demanding Mechanistic Understanding of AI That We Have Never Achieved for Ourselves.* Position paper, working paper, 2026.
7. **P. M. Konrad** (2026). *AutoRed: Measuring the Elicitation Gap via Automated Red-Blue Optimization.* Working paper. Submitted to the Apart Research × Redwood Research AI Control Hackathon, March 2026.

Applied Machine Learning, Software Systems, and Health Informatics

8. **P. M. Konrad**, T. Tanyel, and S. Ayvaz (2025). *Beyond Major Floods: Deep Learning for Detecting Shallow Water Inundation in Agricultural Areas.* Published in *29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025)*, Procedia Computer Science, 270, 301–310.
   ScienceDirect (open access)
9. **P. M. Konrad**, T. L. Adam, R. Terrenzi, and S. Ayvaz (2026). *Architecture Without Architects: How AI*

Coding Agents Shape Software Architecture. Accepted, SAGAI Workshop, IEEE International Conference on Software Architecture (ICSA 2026).

10. T. L. Adam, **P. M. Konrad**, R. Terrenzi, F. G. Lukas, R. Yilmaz, K. Sierszecki, and S. Ayvaz (2026). *CAKE: Cloud Architecture Knowledge Evaluation of Large Language Models*. Accepted, KDA-AI Workshop, IEEE International Conference on Software Architecture (ICSA 2026).
11. R. Terrenzi, **P. M. Konrad**, T. L. Adam, and S. Ayvaz (2026). *Agentic Hybrid Retrieval for Ad Hoc Dataset Search: A Reference Architecture with LLM-Augmented Metadata*. Accepted, SAML Workshop, IEEE International Conference on Software Architecture (ICSA 2026).
12. **P. M. Konrad**, A.-A. Popa, Y. Sabzehmeidani, L. Zhong, E. A. Liehn, and S. Ayvaz (2025). *Challenges in Deep Learning-Based Small Organ Segmentation: A Benchmarking Perspective for Medical Research with Limited Datasets*. Under revision, *Biomedical Signal Processing and Control*; preprint at arXiv:2509.05892.
    arXiv preprint
13. **P. M. Konrad**, Y. Sabzehmeidani, A.-A. Popa, and S. Ayvaz (2025). *Machine Learning in Gastrointestinal Tract Imaging: A Comprehensive Review of Techniques and Applications*. Under review, *Journal of Imaging Informatics in Medicine*.
14. **P. M. Konrad**, C. H. Kunstmann-Olsen, J. Fiutowski, and S. Ayvaz (2025). *Non-Destructive Prediction of Fruit Ripeness and Firmness Using Hyperspectral Imaging and Lightweight Machine Learning Models*. Under review, *Computers and Electronics in Agriculture*.
15. **P. M. Konrad** (2026). *The AI Productivity Measurement Problem: Construct Mismatches Explain Why Coding Tool Studies Disagree*. Preprint, 2026.
16. **P. M. Konrad** and T. L. Adam (2026). *The Trader's Trinity: Forecasting Models, RL Agents, and LLM Judges for Day-Ahead Markets*. **BSc thesis**, University of Southern Denmark. Work in progress.

Research Experience

Independent Research, Mechanistic Interpretability Sep 2025 – Present
Self-directed and self-funded Remote

- Solo working papers on sandbagging circuits, differential dishonesty, moral-framing geometry, fixed-point limits of self-prediction, the PUB benchmark proposal, and automated red-blue optimisation (entries 2–7 above).
- Open-weight models in the 0.5B–8B range (Gemma 2/3, Qwen 2.5, LLaMA 3.1, Mistral) instrumented end-to-end on a single Apple Silicon MacBook with TransformerLens, nnsight, SAE-Lens, and custom hook libraries.
- Full lifecycle owned independently: problem formulation, experimental design, implementation, statistical validation (bootstrap confidence intervals, permutation testing, walk-forward validation), manuscript writing, and venue selection.

DataVISards, Hong Kong University of Science and Technology (HKUST) Jan 2026 – Present
Research Collaborator Hong Kong SAR, China

- Collaborating with the DataVISards research group on machine learning and data visualisation projects building on coursework from the COMP6411D postgraduate course.

Data and Intelligence Lab, SDU Sep 2024 – Present
Research Collaborator (Jan 2026 – Present); *Research Assistant* (Sep 2024 – Dec 2025) Sønderborg, Denmark

- Independently initiate and drive applied ML research projects under Assoc. Prof. Serkan Ayvaz, spanning remote sensing, medical imaging, hyperspectral imaging, and software architecture.
- Co-author of one published KES 2025 paper and seven additional manuscripts at varying stages (accepted, under revision, under review, preprint).
- Contributions to grant-proposal preparation and multimodal dataset management across the lab's research portfolio.

Academic Service

Teaching Assistant, Artificial Intelligence (BSc, SDU) Jan 2026 – Jun 2026

- Designing hands-on lab assignments, supporting exercise sessions, and contributing to curriculum development for the BSc-level Artificial Intelligence course.

Educational Committee, Software Engineering Programme, SDU Jun 2025 – Present

Student Member — Sønderborg, Denmark

- Represent BSc Software Engineering students in curriculum and quality-assurance discussions, providing feedback on course content, assessment, and study environment.

Professional Experience

SaturoLabs — Jan 2026 – Present
Founder — Denmark

- Solo umbrella for product experiments at the intersection of AI and human flourishing.
- Flagship: **DreamBear**, an AI bedtime-story app for neurodivergent children aged 3–10, where ADHD energy, autism-related focus, and dyslexic creativity become heroic superpowers in personalised narratives. iOS 16+, freemium, COPPA-compliant, no ads or data sharing. Built on the Anthropic Claude API, Next.js, and ElevenLabs voice synthesis.
- Additional product surfaces: claudeboyz.com, getproofz.com.

Tutora ApS — Jun 2024 – Nov 2024
CTO and Co-Founder — Sønderborg, Denmark

- Led end-to-end development of the company's websites and core web application.
- Implemented the Shape Up product development framework to streamline technical execution.

Yeager GmbH — Nov 2022 – Jun 2024
Co-CEO and Co-Founder — Remote

- Co-founded the company and led development of *stabil.ai*, an AI-powered mobile app for personalised powerlifting training.
- Engineered algorithms to personalise training plans using MRV/MEV principles and dynamic real-time feedback.
- Owned the full product lifecycle from UX/UI to full-stack implementation under lean-startup principles.

Bundeswehr (German Armed Forces) — Oct 2017 – Sep 2021
Staff Duty Soldier — Glücksburg, Germany

- Led a small HR team responsible for the administration of more than 600 soldiers in a high-stakes naval command environment.
- Decorated twice for sustained excellence and independent handling of complex tasks.

Skills

Mechanistic Interpretability: TransformerLens, nnsight, SAE-Lens, custom hook libraries, activation patching, linear probing, logit lens, ablation studies, residual-stream analysis, model-organism construction
Deep Learning: PyTorch (MPS), Hugging Face Transformers, accelerate, peft, trl, datasets
Open-Weight Models: Gemma 2/3, Qwen 2.5, LLaMA 3.1, Mistral (typically 0.5B–8B, MacBook-reproducible)
Experimentation & Statistical Validation: Weights & Biases, Optuna, fixed seeds, walk-forward validation, bootstrap confidence intervals, permutation testing
Data & Analysis: NumPy, Pandas, scikit-learn, scipy, Matplotlib, Seaborn
Programming Languages: Python, JavaScript / TypeScript, SQL
Product Engineering: Next.js, React, React Native, Node.js, Vercel, Anthropic Claude API, ElevenLabs, Docker, Google Cloud Platform

Languages

German (Native or bilingual) **Thai** (Native or bilingual) **English** (Professional working) **Danish** (Elementary)

Honors and Awards

Top-10 Placement, Danish National Championship in AI (DM i AI) — 2025
National AI competition, individual competitor

- Secured a top-10 placement competing solo against more than 50 teams in Denmark's largest AI competition, organised by the Danish Society for Artificial Intelligence.

Top-10 Placement, Danish National Championship in AI (DM i AI) — 2024
National AI competition, team competitor

- Top-10 placement in Denmark's national AI competition for students and professionals, organised by the Danish Society for Artificial Intelligence.

1st Place, SDU Case Competition, SDU Sønderborg 2023
University case competition on sustainability

- Awarded 1st place in a 48-hour competition focused on sustainability with real-world cases from Danfoss, Linak, and partner companies.

Formal Recognition for Exemplary Service, Bundeswehr (German Armed Forces) 2021

- Formally commended for setting a benchmark in dedication and professionalism for enlisted personnel.

Performance Bonus for Outstanding Achievement, Bundeswehr (German Armed Forces) 2021

- Awarded a significant financial bonus for sustained excellence and independent handling of complex tasks.

Certifications

Venture Capital Explorer Programme, Accelerace & BII 2026
Intensive 4-day program covering VC investment process, founder sourcing, startup assessment, venture risk evaluation, investment committee operations, exits, and case-study simulations.

Machine Learning, Stanford Online 2024
Foundational C# with Microsoft 2024
Google UX Design Professional, Google 2023
React Native Course 2023
Full-Stack Engineer Career Path, Codecademy 2023
Certified Specialist for Real Estate Loan Brokerage (IHK) 2022