

Jeremias Lino Ferrao

AI RESEARCHER — MODEL UNDERSTANDING, EVALUATION, AND CONTROL

Groningen, Netherlands

j.lino.ferrao@gmail.com | [Jazhyc](#) | [LinkedIn](#) | [Google Scholar](#)

Education

University of Groningen

Groningen, Netherlands

• MSc in Artificial Intelligence

September 2024 - July 2026

- GPA - 9.2, Summa Cum Laude, Thesis supervised by Yftah Ziser
- Machine Learning Specialization: Deep Learning, Reinforcement Learning, Natural Language Processing

• BSc in Artificial Intelligence

September 2021 - July 2024

- GPA - 8.7, Cum Laude, Thesis published at Northern Lights Deep Learning Conference 2025

Publications

Intent-Aware Training Improves LLM Safety Classification Across Training Regimes

Budapest, Hungary

First Author, Accepted to Findings of EMNLP 2026

[Link](#)

- Coordinated human annotation of 1,724 intent-label pairs targeting adversarial and hard-to-classify safety prompts.
- Trained intent-aware SFT and distillation models across 144 conditions, surpassing GPT-5.4 and dedicated classifiers on 10–100x less data.
- Showed that our intent-aware models dominate the latency-F1 Pareto frontier over prior guardrails at a fraction of their inference latency.

The Anatomy of Alignment: Decomposing Preference Optimization by Steering Sparse Features

San Diego, United States

First Author, Spotlight at NeurIPS Mechanistic Interpretability workshop 2025

[Link](#)

- Introduced FSRL, a transparent alignment framework that steers models via interpretable SAE features.
- Demonstrated FSRL is comparable to standard RLHF methods while providing mechanistic insight into the alignment process.
- Found that the alignment process learned to prioritize stylistic features (e.g., formatting) over high-level alignment concepts.

Self-Ablating Transformers: More Interpretability, Less Sparsity

Singapore

First Author, Poster presentation at Building Trust workshop at ICLR 2025

[Link](#)

- Developed a novel transformer architecture that enhances interpretability by dynamically ablating neurons and attention units during training.
- Achieved a 4x reduction in circuit size, resulting in more understandable model behaviour with a minor cost to language modelling performance.
- Models utilize standard transformer architecture during inference, maintaining full interoperability with existing frameworks.

World Model Agents with Change-Based Intrinsic Motivation

Tromsø, Norway

First Author, Oral presentation at NLDL 2025

[Link](#)

- Adapted CBET intrinsic reward technique, originally for RL model-free algorithms, to the state-of-the-art DreamerV3 world model algorithm.
- CBET enhances DreamerV3's returns by $\approx 5\%$ in the complex Crafter environment, showing promise for handling sparse rewards.
- CBET found to be detrimental in simpler Minigrid environment suggesting misalignment between extrinsic and intrinsic rewards for DreamerV3.

Experience

MATS 12.0 – Neel Nanda's Exploration Phase

Remote

Scholar

September 2026 - October 2026

- Investigating training efficiency improvements for action-only monitors used in AI control.
- Exploring mechanistic interpretability methods to improve monitor robustness to prompt injection attacks.

Artificial Intelligence Safety Initiative, Groningen

Groningen, Netherlands

Technical Alignment Lead

October 2023 - June 2026

- Supervised a team of 3 junior researchers on a model organism project investigating evaluation awareness in LLMs at tractable scales.
- Led and managed the technical discussion group, focusing on mechanistic interpretability and alignment research.
- Coordinated the technical track of the AI Safety Fundamentals course comprising 2 facilitators and 17 students across 8 weeks.

Apart Research

Remote

Research Fellow

September 2024 - March 2026

- Led and first-authored two research projects in mechanistic interpretability, yielding submissions to workshops at top-tier conferences.
- Managed 100+ experiments for up to 9B parameter LLM training and evaluation across both HPC clusters and public cloud platforms.
- Architected reproducible research pipelines with robust experiment tracking (Wandb) and configuration management (Hydra, UV).

EleutherAI – Summer of AI Research Program

Remote

Research Mentee

August 2025 - October 2025

- Conducted a benchmark of open source models for SAE feature explanation, using vLLM to enable high-throughput evaluation.
- Established a new Pareto frontier, identifying models that achieve SOTA-level quality with a 3x inference speedup.
- Evaluated contrastive examples for feature explanation, finding they increase scoring rigor but do not improve generation quality.

Skills

Frameworks	Pytorch, Hugging Face, Wandb, UV, Hydra, vLLM, Scikit-learn, FAISS, Pandas, Numpy, Matplotlib, NLTK
Miscellaneous	Agentic Software Development (Codex, Claude Code), SLURM/HPC, Linux, Git, Shell, Docker, Debugging, LaTeX
Soft	Scientific Writing, Reviewing, Teamwork, Leadership, Problem-solving, Documentation, Presentation

Awards

Winner	Aletheia's Quest - 1st Public Leaderboard, 2nd Black-Box Track & Scalability Award	August 2026
1st Place	NS-Gym AAMAS 2026 Competition - Unsupervised Environment Design in Non-Stationary Settings	May 2026
3rd Place	AI Safety x Physics Grand Challenge - Developmental Interpretability in Reasoning Models	July 2025
1st Place	Best Student Paper for Advanced Topics in NLP course: Attribution in LLMs	April 2025
3rd Place	AIMZ Energy Prediction Kaggle Competition - Ensemble of SVM, MLP, GBM and Stacked LR	December 2024
1st Place	Apart X Goodfire Reprogramming AI models Hackathon - Inference Time Interpretable RL in LLMs	November 2024
2nd Place	Apart AI Security Evaluation Research Hackathon - LLM Cybersecurity Persistence Benchmark	May 2024

Academic Service

Reviewer	ICML 2026 & NeurIPS 2025 Mechanistic Interpretability Workshops, NLDL 2026, ICLR 2025 Building Trust Workshop
-----------------	---

Notable Projects

Rethinking the Monitoring Interface for AI Control (Successful Neel Nanda Application) September 2026 | [Link](#)

- Identified a monitoring bottleneck in prior work: ordinal scores conflated violation severity with model confidence.
- On six OOD datasets, binary-logit scoring let larger base models outperform trained frontier monitors at lower inference cost.
- Soft distillation further raised the low-cost OOD frontier: 0.7823 vs. 0.6840 mean pAUROC@20 at 27% lower inference cost.

Scalable LLM Deception Detection via Soft Distillation (Aletheia's Quest) August 2026 | [Link](#)

- Found that soft teacher targets preserved ranking information more effectively than hard binary decisions for AUROC.
- Soft-distilled Kimi K3 into a rank-16 Qwen3.5-9B LoRA trained on 13,149 examples for efficient black-box inference.
- Achieved 0.9661 mean AUROC with 18x faster inference than EleutherAI's runner-up submission on the public leaderboard.

Unsupervised Environment Design for Non-Stationary Reinforcement Learning (NS-Gym) May 2026 | [Link](#)

- Reframed non-stationary RL as a UED problem by treating environment configurations as levels, enabling PLR/ACCEL to act as a curriculum.
- Built a high-throughput asynchronous PPO pipeline, finding that off-policy corrections hurt performance in non-stationary environments.
- Outperformed planning and continual learning methods from competitors in both returns and inference latency.

Passing the Test, Breaking the System: Latent Failures in Software Engineering Agents March 2026 | [Link](#)

- Designed a 36-task benchmark to evaluate agents on latent failures—defects that pass functional tests but cause subsequent system harm.
- Evaluated 11 models, demonstrating a gap where agents achieved high primary task success (78-94%) but low latent safety success (18-36%).
- Showed that pre-task environment scouting and open-ended prompting improved latent safety without reducing primary task completion.

Predicting Developmental Milestones in RL with Local Learning Coefficients July 2025 | [Link](#)

- Investigated LLM capability emergence in GRPO by framing it as a phase transition, using Singular Learning Theory for analysis.
- Proposed using reward function components as direct observables for these transitions, simplifying developmental interpretability.
- Demonstrated a strong correlation between spikes in the Local Learning Coefficient and shifts in model behavior during RL training.

From Steering Vectors to Conceptors: Compositional Affine Activation Steering for LLMs May 2025 | [Link](#)

- Developed and managed the experimental infrastructure to validate a new, theoretically-justified activation steering method.
- Conducted experiments that demonstrated improved steering performance over compared baselines.

Examining Step and Token Level Attribution in Multilingual Chain-of-Thought Reasoning April 2025 | [Link](#)

- Investigated the interpretability and faithfulness of multilingual Chain-of-Thought reasoning in LLMs using feature attribution techniques.
- Implemented step-level attribution for Qwen 2.5 1.5B Instruct and its R1 distilled variant using structured generation on the MGSM benchmark.
- Analysis revealed base model's final-step focus contrasted with distilled model's more holistic reasoning, highlighting variations in faithfulness.