Matteo He

matteohe.research@gmail.com | matteohe.com | github.com/hematteo

Selected Research

Sparse Readout Prism: Explaining Logit-Lens Scores in Features Instead of Tokens
Matteo He, William F. Shen, Xinchi Qiu, Nicholas D. Lane. 2026.
First author; preprint under review | arXiv:2609.01936 | Code | Dictionaries

- Showed that a lens's fitting corpus can change its reported language for identical hidden states in Qwen3.5-9B. Tested Jacobian lenses fitted on English and Chinese, with controls using English and German and ridge lenses.
- Developed SRP to decompose lens scores into shared feature contributions and explicit residuals by fitting sparse autoencoders to output weights. Tested predicted score changes by ablating features in the readout.
- Improved reconstruction coverage for logit differences by **8.9 to 17.3 percentage points** over the strongest of six geometric baselines across six readouts without softcaps. Coverage requires the correct sign and less than 50% relative error.

Learning to Read Out: Unembedding Dynamics in Language Model Pretraining
Matteo He, William F. Shen, Alex Iacob, Andrej Jovanovic, Xinchi Qiu, Nicholas D. Lane. 2026.
First author; accepted to NeurIPS 2026 | Code

- Showed why output scores can misdate capability acquisition during training: both hidden representations and the learned readout change. Designed linear probes and readout swaps to separate their contributions.
- Found subject–verb agreement linearly decodable in Pythia-6.9B several hundred steps before the model's readout expressed it. Gaps persisted at convergence on harder contrasts involving relative clauses.
- Developed trajectory crosscoders to track sparse readout features during pretraining in Pythia and OLMo-2-7B. Studied early readout reorganization through dictionary dynamics, lexical probes, and readout swaps.
- In control runs with 31M parameters and one seed per condition, each 4× increase in readout learning rate halved the reorganization step. Final validation losses agreed within 0.06 nats.
- Localized agreement scores to eight readout features. With Pythia-1B hidden states fixed at step 1000, ablating these features from the reconstructed terminal readout reduced agreement accuracy on held out examples from 90.0% to 48.7%; matched controls stayed at 90.0%.

Automating the Interpretability of Attention Heads in LLMs
BSc Computer Science dissertation, University of St Andrews 2024

- Built a PyTorch and Transformers pipeline to generate and simulate explanations of GPT-2 attention heads. Compared predicted and observed attention using distributional errors and errors for individual tokens.

Education

University of Cambridge Oct 2025 - Jun 2026
MPhil in Advanced Computer Science, Distinction; AI Alignment Fellowship

University of St Andrews Sep 2021 - Jun 2024
BSc (Hons) Computer Science and Mathematics, First Class Honours
Average **18/20**; Dean's List all three years.

Industry Experience

Amazon, Alexa AI Jun 2023 - Sep 2023
Software Development Engineer Intern, ML Evaluation Infrastructure
Built Python and AWS evaluation pipelines processing millions of utterances daily for **20+ Applied Scientists**.

Selected Awards

Top Student Medal: Highest academic achievement in the Direct Entry Computer Science cohort.
GRE: 340/340. **Challenge wins:** OxfordHack 2022 (Oxbotica); HackTheBurgh VIII (GitHub).

Technical Skills

Research methods: Sparse coding, linear probes, readout swaps, ablations, LLM evaluation.
Tools: Python, PyTorch, Transformers, TransformerLens, JAX, GPU training, Git/Linux, AWS.