AI SAFETY · ALIGNMENT RESEARCH

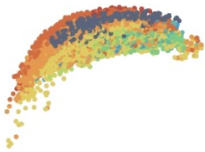
Ian Rios-Sialer

interpretability · evals · llm bias

Mechanistic interpretability of bias, preference, and diversity in language models.

Time horizon, localized in the residual stream (Qwen3-4B)

long vs short horizon



seconds → centuries

AT A GLANCE

Honorable Mention · Apart 2026

ex-Niantic · Research Engineer

TAIS · EurIPS · ICML

TALK TO ME ABOUT

Applied interpretability · functional concepts · bias evals.

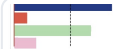
Scan for papers, code & CV

unrulyabstractions.com

SELECTED WORK — first author

Temporal Preference Concepts and their Functions in a Large Language Model
arXiv:2606.05194 · time-horizon geometry in the residual stream

The Homogenization Problem in LLMs: Towards Meaningful Diversity in AI Safety
arXiv:2601.06116 · measuring homogenization & output diversity



gender bias, Claude 3.5 Haiku

MORE

- **Scale Shapes Bias in Spanish-Prompted Open-Weight LLMs**
  Honorable Mention · Global South AI Safety Hackathon · Apart × Schmidt
- **Towards Local Bias Mitigation: How LLMs Treat LGBTQ+ Users Differently, Without Explicit Markers**
  working paper · unrulyabstractions.com
- **Category-Theoretic Wanderings into Interpretability**
  essay · unrulyabstractions.com
- **Which Circuit is it?**
  LessWrong sequence · toy-model interpretability

PRESENTED AT

Technical AI Safety Conf. (TAIS) · EurIPS · ICML

COLLABORATIONS

Apart · SPAR · AI Safety Camp · Queer in AI · Groundless · Magic Fund

BEFORE AI SAFETY · 8+ YRS SHIPPING ML

Niantic · Pokémon Go AR · 6D.AI · Civil Maps

BSE Mechanical Engineering · U. Michigan · Magna Cum Laude '17

+1 734 546 7145 **iansebas@umich.edu**

linkedin.com/in/iansebas · iansebas.github.io