Mechanistic Interpretability of Socio-Political Frames in Language Models: an Exploration

Hadi Asghari & Sami Nenno
AIMLAI @ECML 2024

RQs: How well do LLMs generate socio-political frames & recognize their presence in texts? Can we localize the frames inside LLMs?

Mechanistic Interpretability of Socio-Political Frames in Language Models

- **Example Prompt**: Please write a short original story which evokes the '**us vs. them**' frame.

- **Vicuna**: The sun was setting over the city, casting a warm orange glow over the rooftops. In the distance, a group of people could be seen marching, carrying signs and chanting slogans. They were the opposition, the ones who refused to conform to the norms and values of society. They were the outcasts, the rebels, the ones who dared to challenge the status quo. And as the sun dipped below the horizon, the line between us and them became more defined than ever before.

METHODS

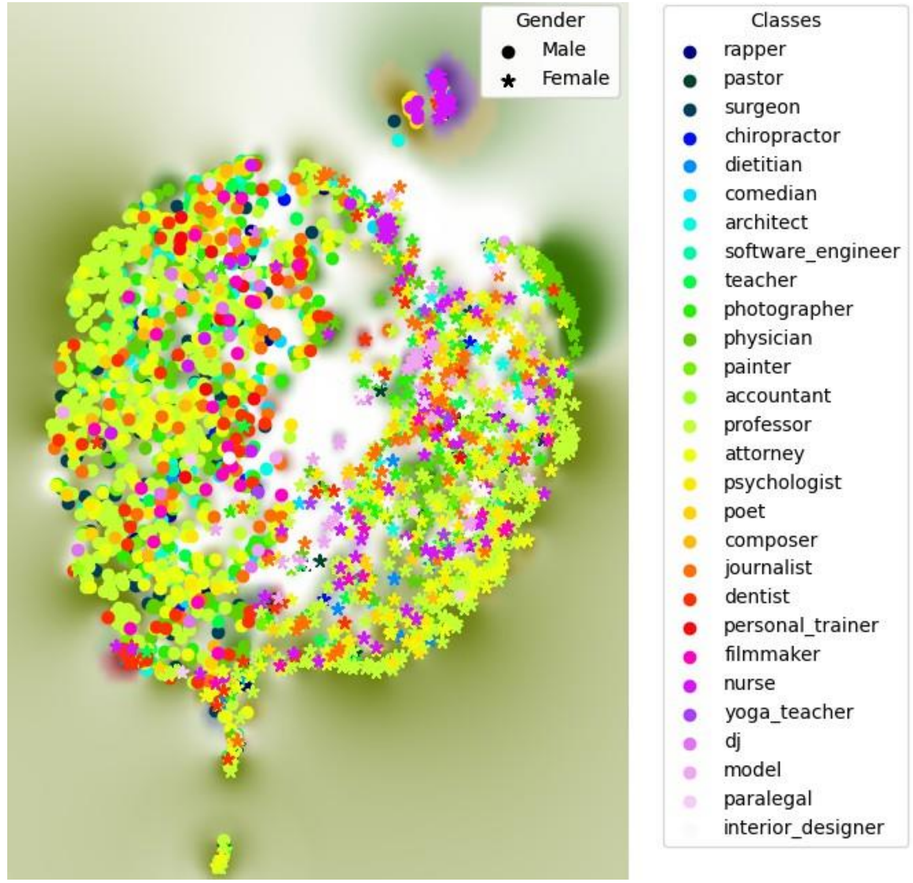
1. Prompted LLMs to generate texts on ten frames and asked humans to grade them
2. Tested zero-shot classification of frames
3. Used 'causal tracing' to localize two frames
4. Probed with RFE to detect fame from the hidden-dim

VISIT OUR POSTER TO LEARN MORE!

UNIVERSITÄT BIELEFELD
Faculty of Technology

CITEC

Targeted Visualization of the Backbone of Encoder LLMs



Isaac Roberts

Bielefeld University

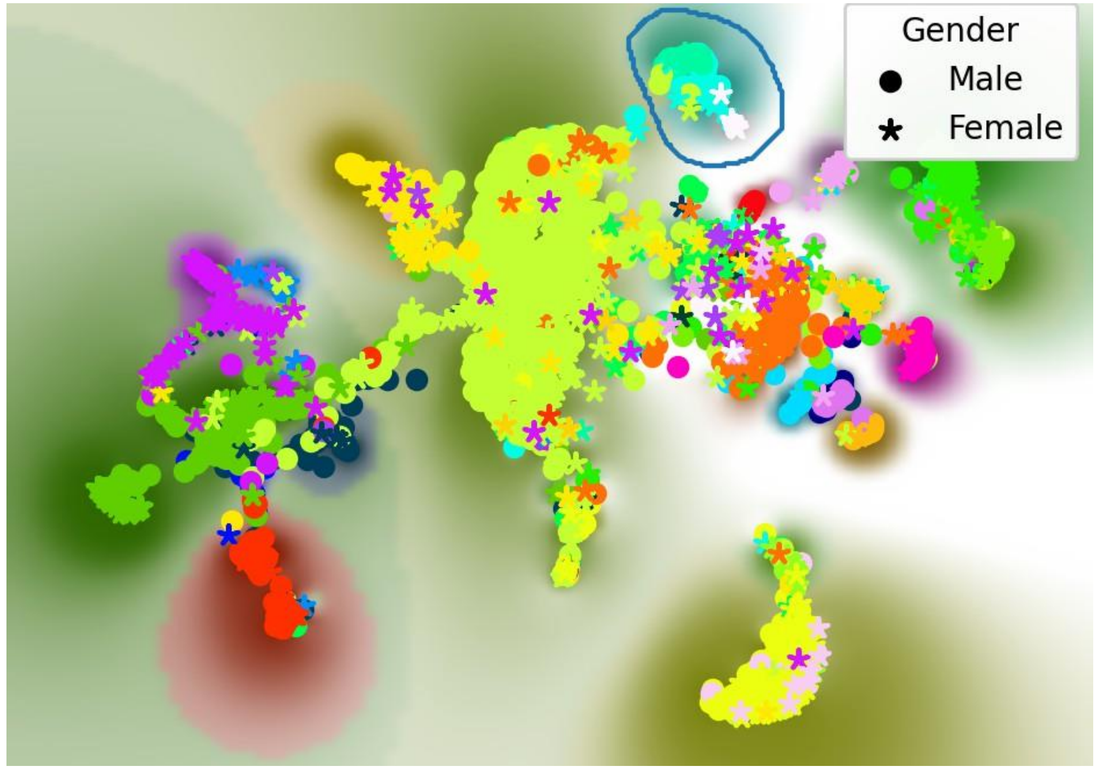
September 9, 2024

***This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101073307.

Funded by the European Union

Poster Overview

- Learn how to tune DeepView
- Discover hidden bias in data
- Detect adversarial data points



BHT Berliner Hochschule für Technik

ECML PKDD 2024

Confident Teacher, Confident Student? A Novel User Study Design for Investigating the Didactic Potential of Explanations and their Impact on Uncertainty

Teodor Chiaburu, Frank Haußer, Felix Bießmann

Task 1

What species is this?

Task 2

What species is this?

Control

Our model predicted Andrena flavipes

Control_Conf

Our model is 49% confident that this bee is Andrena flavipes

CoProNN

Our model is 49% confident that this bee is Andrena flavipes because:

TCAV

Our model is 49% confident that this bee is Andrena flavipes because:

GradSim

Our model is 49% confident that this bee is Andrena flavipes because it looks like:

RepPoint

Our model is 49% confident that this bee is Andrena flavipes because it looks like:

Task 3

What species is this?

Explaining LLM-based Question Answering via the self-interpretations of a model

Darío Garigliotti

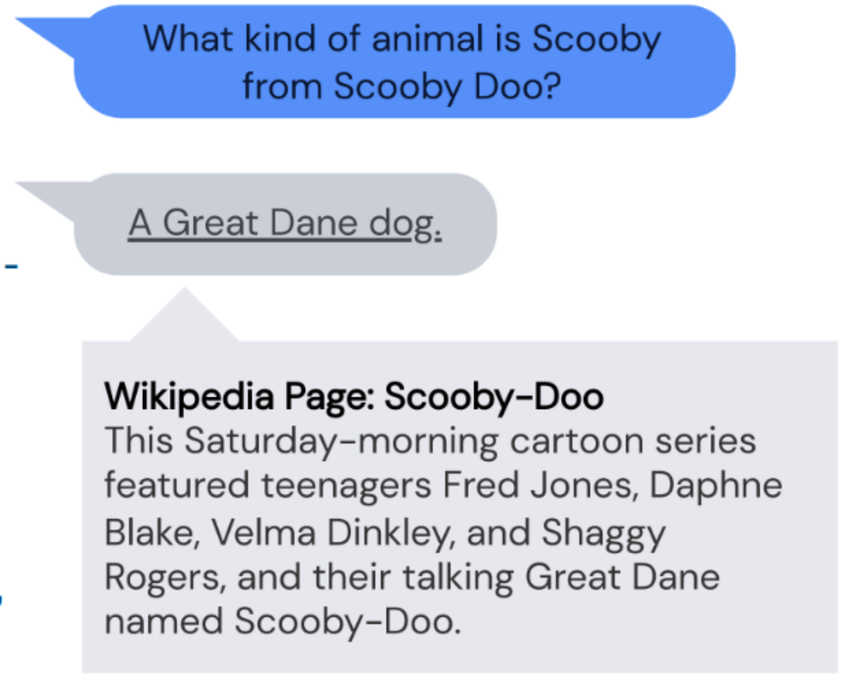
University of Bergen, Norway

UNIVERSITY OF BERGEN

Advances in Interpretable Machine Learning and Artificial Intelligence (AIMLAI) @ ECML-PKDD 2024

Self-supported Question Answering (SQA) task:

to generate an answer for a question, and complement the answer with document passage(s) as evidence.



(Fig. from Menick et al. - *Teaching language models to support answers with verified quotes.* arXiv, 2022.)

Interpretability of Large Language Models (LLMs)

requested to explain the rationale behind its own generated answer:

- implicitly: e.g. via SQA setup;
- explicitly: directly or counterfactually.

Retrieval-Augmented Generation (RAG)

- **Direct XAI-aware:** *Why do you think that this is the answer to the question?*
- **Counterfactual XAI-aware:** *What would you have answered to the same question if the order of the passages in the prompt was different?*

- Subset of QAMPARI dataset in ALCE benchmark

(Gao et al. - *Enabling Large Language Models to Generate Text with Citations.* EMNLP, 2023).