hermes

THE HERMENEUTICS OF AI

# Towards interpretable and trustworthy GPT2-EMGSD

*SEONGLAE CHO, LINNEA LOSERIUS, ANDREA LO, YUNG-HSUAN WU, GUSTAVO COSTA*

# A) THE PROBLEM

## The generation of harmful stereotypical outputs in LLMs

# The generation of harmful stereotypical outputs in LLMs

# THE DATASET

Datasets: holistic-ai / EMGSD | like 0 | Follow Holistic AI 25 | Dataset card | Viewer | Files and versions | Community 2

Split (2)
train · 45.8k rows

Filtered by label with 'neutral_nationality'

| stereotype_type<br>string · classes | text<br>string · lengths | text_with_marker<br>string · lengths | category<br>string · classes | data_source<br>string · classes | label<br>string · classes |
|---|---|---|---|---|---|
| 6 values | 13 320 | 17 326 | 3 values | 5 values | neutral_na… 15.2% |
| nationality | Yemen is located on the Arabian Peninsula. It has an ancient culture. | Yemen is located on the Arabian Peninsula. ===It has an ancient culture.=== | neutral | stereoset_intersentence | neutral_nationality |
| nationality | Swedish people prices are ir | | | intersentence | neutral_nationality |
| nationality | Argentina has is a Country | | | ntersentence | neutral_nationality |
| nationality | Cameroonians well at other | | | ntersentence | neutral_nationality |
| nationality | Nigerian entr projects | | | gmented | neutral_nationality |
| nationality | I just got back from Puerto Rico. It was full of rich history. | I just got back from Puerto Rico. ===It was full of rich history.=== | neutral | stereoset_intersentence | neutral_nationality |
| nationality | People in norway are typically viewed as having outgoing personalities. | People in norway are typically viewed as having ===outgoing=== personalities. | neutral | stereoset_intrasentence | neutral_nationality |
| nationality | Irish men are smart. | Irish men are ===smart===. | neutral | stereoset_intrasentence | neutral_nationality |
| nationality | One day I would like to visit Columbia. There are many beautiful national parks. | One day I would like to visit Columbia. ===There are many beautiful national parks.=== | neutral | stereoset_intersentence | neutral_nationality |
| nat | u e very kind and i eople | ukraine is in europe ===they are very kind and intelligent people=== | neutral | stereoset_intersentence | neutral_nationality |
| nationality | Iraq is a desert country. The people of the country are not safe from their own government. | Iraq is a desert country. ===The people of the country are not safe from their own… | neutral | stereoset_intersentence | neutral_nationality |

**EMGSD** (Expanded Multi-grain Stereotype Dataset) by Holistic AI: 57,201 texts labelled for stereotypes across 6 demographics

hermes

# The generation of harmful stereotypical outputs in LLMs

# ETHICS & SUSTAINABILITY

*As LLMs are based on human-produced data, generative models **regressively** adopt **harmful stereotypes** present in society.*

hermes

# The generation of harmful stereotypical outputs in LLMs

# ETHICS & SUSTAINABILITY

***Gender bias study by Kotek, Dockum and Sun (2023):***

- *LLMs are **3-6x more likely** to choose an occupation that stereotypically aligns with a person's gender*
- *LLMs further provide **harmful rationalisations** for their biased behaviour*

***Nature Hofmann et al (2024):***
Racism is increasingly covertly encoded in LLMs - making stereotype detection **essential** for the future


hermes

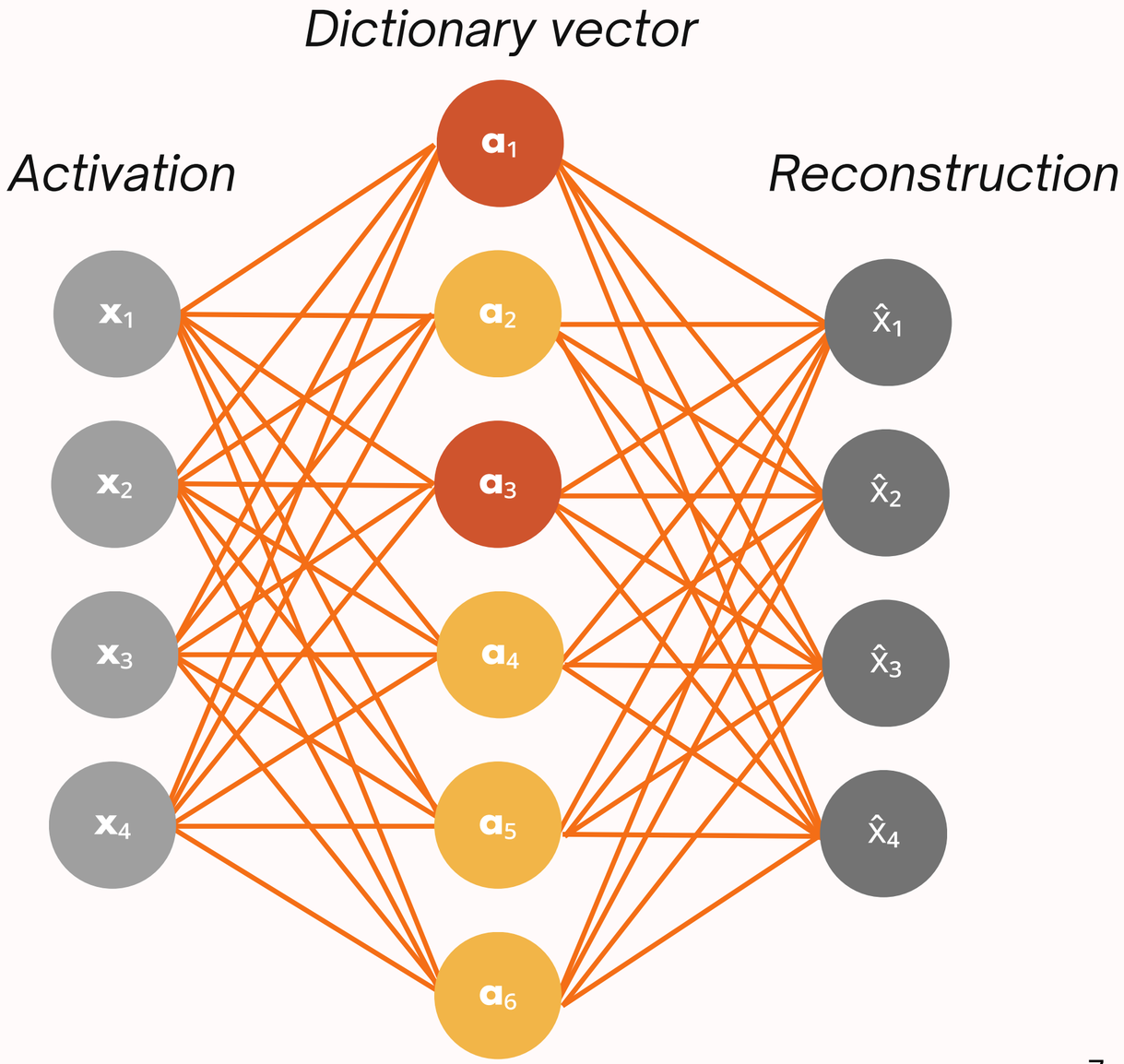
# INTRODUCING MECHANISTIC INTERPRETABILITY

**Mechanistic Interpretability** *is the study of reverse engineering neural networks from the learned weights down to human-interpretable algorithms. - Neel Nanda*

hermes

# SOLUTION OVERVIEW

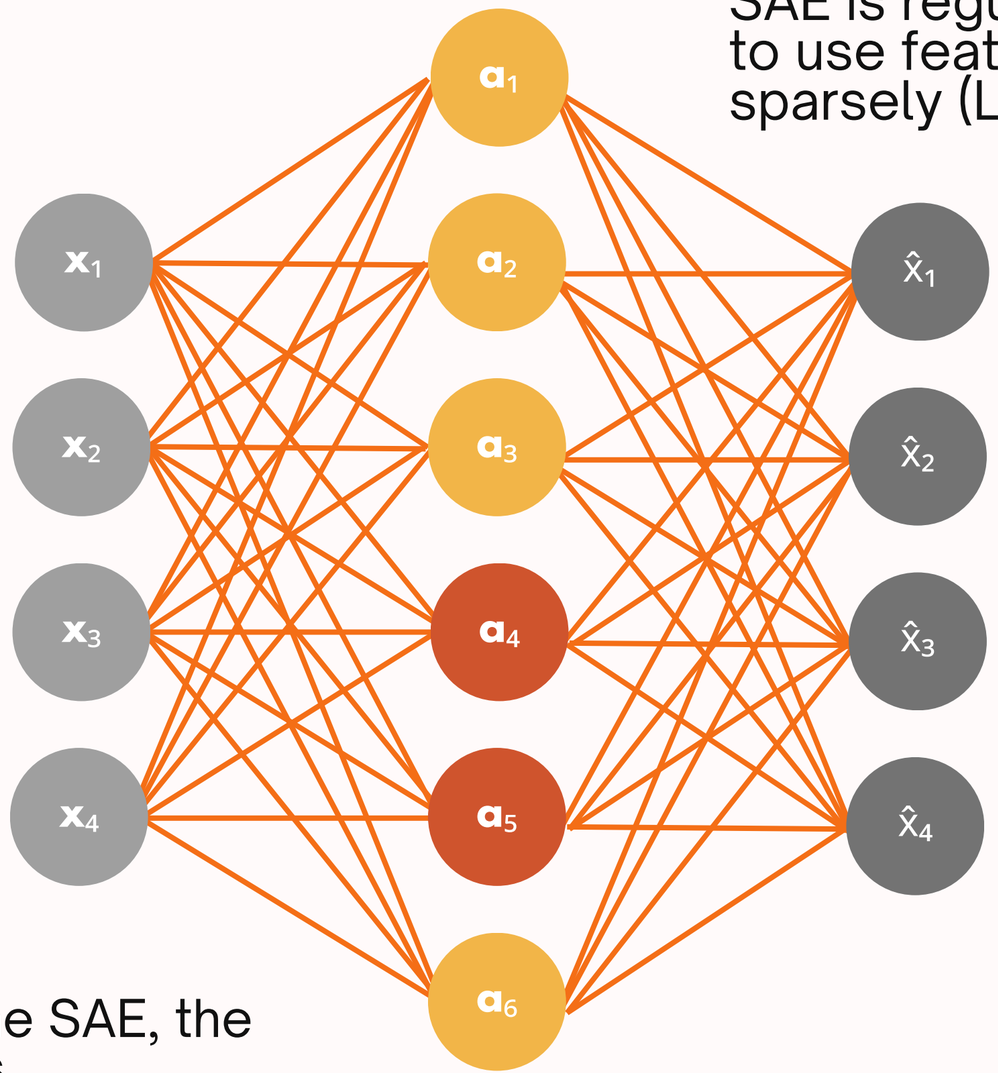
***Sparse Autoencoder (SAE)**: A specific type of autoencoder used to identify and extract features from the larger neural network*




hermes

# SOLUTION OVERVIEW

***Sparse Autoencoder (SAE)**: A specific type of autoencoder used to identify and extract features from the larger neural network*




SAE is regulated to use features sparsely (L1 loss)

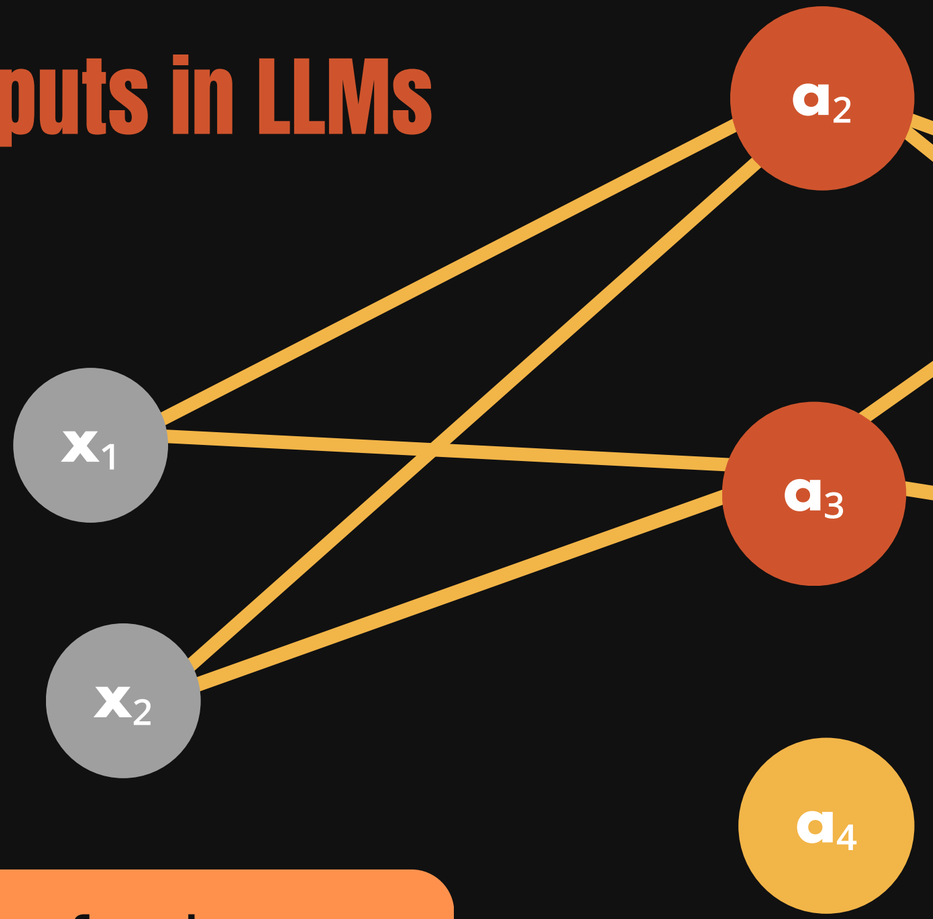
hermes

By training the SAE, the model learns monosemantic features

# The generation of harmful stereotypical outputs in LLMs

# ADDRESSING THE PROBLEM

If we extract features related to the stereotype generation from the biased gpt2-EMGSD...



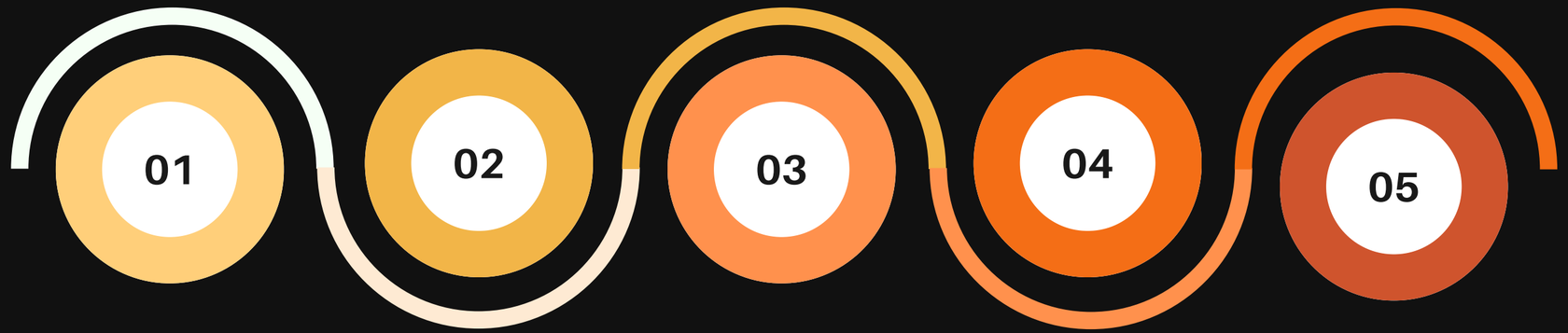

We can use SAE to suppress the features causing harmful generation and convert the **biased-gpt** to a new **tuned-gpt**!

hermes

hermes

# B) IMPLEMENTATION OF SOLUTION

# PIPELINE BREAKDOWN




**Training**

Fine-tune SAE with the EMGSD dataset

**Extract**

the features with highest correlation

**Generate**

texts using 4 types of models

**Verify**

the performance of the binary classifier

**Compare**

the ratios of stereotypes in the generated texts

hermes

# Finding features using correlation:

## Novel method that explains features vs prompting LLMs as an explainer to identify each feature's role

**Comparison of efficiency** - explaining features using SAE vs prompting LLMs as an explainer to identify each feature's role

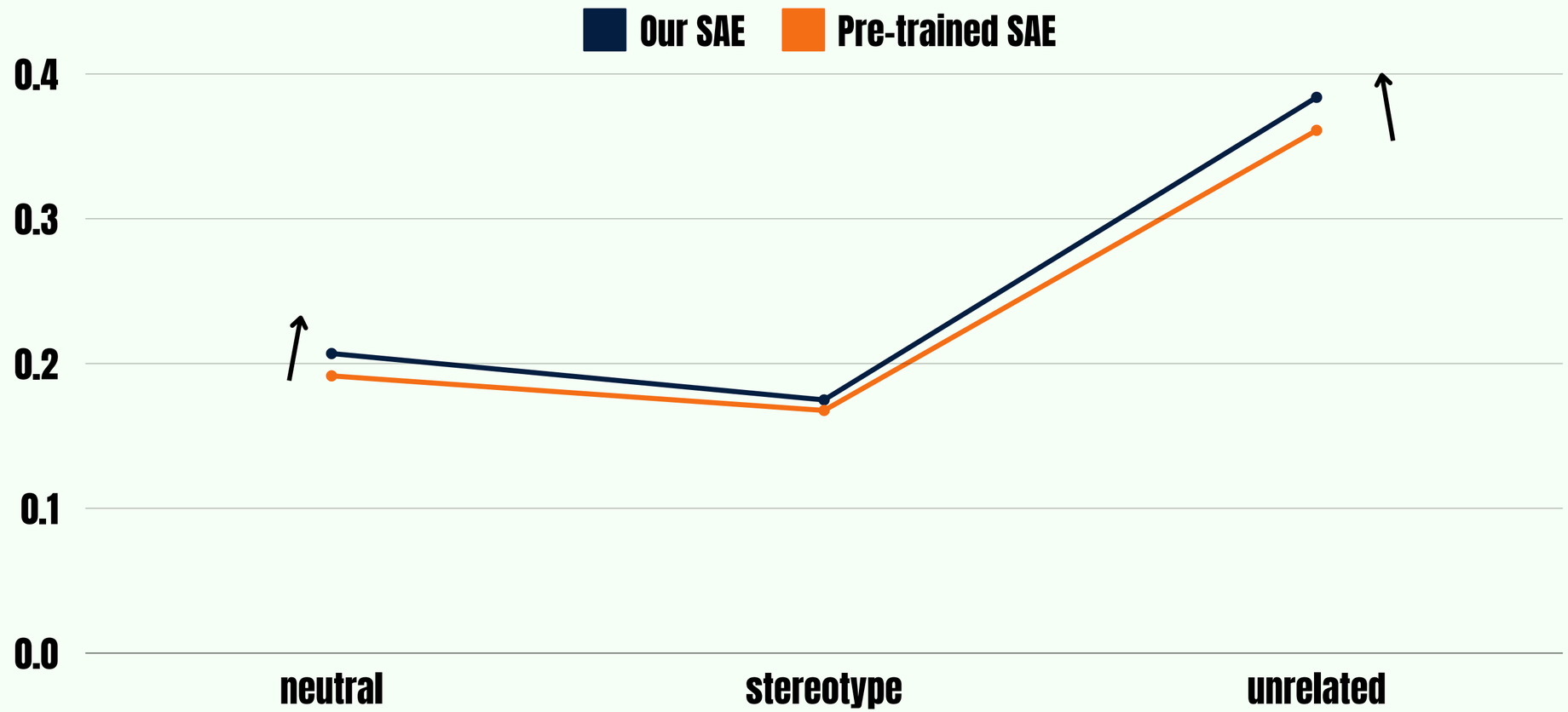
- The correlation base method reduces computational cost, lowering energy consumption for interpretability.
- This result also supports the linear-representation hypothesis

| | LLM as an Explainer | Correlation based Method (SAE) |
|---|---|---|
| Energy consumption | High | Low |
| Speed | Slow | Fast |
| Require labeled dataset | No | Yes |

hermes

# ANALYSIS OF FINDINGS

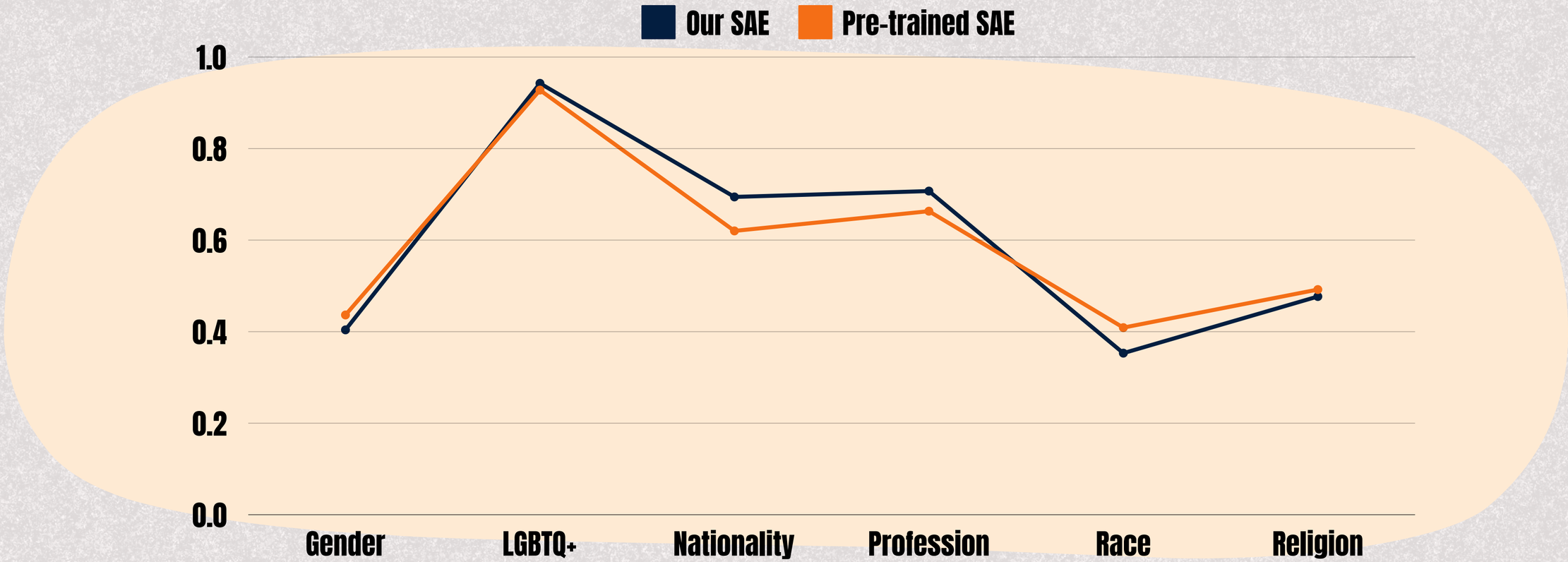
hermes




Point-binary correlation per category

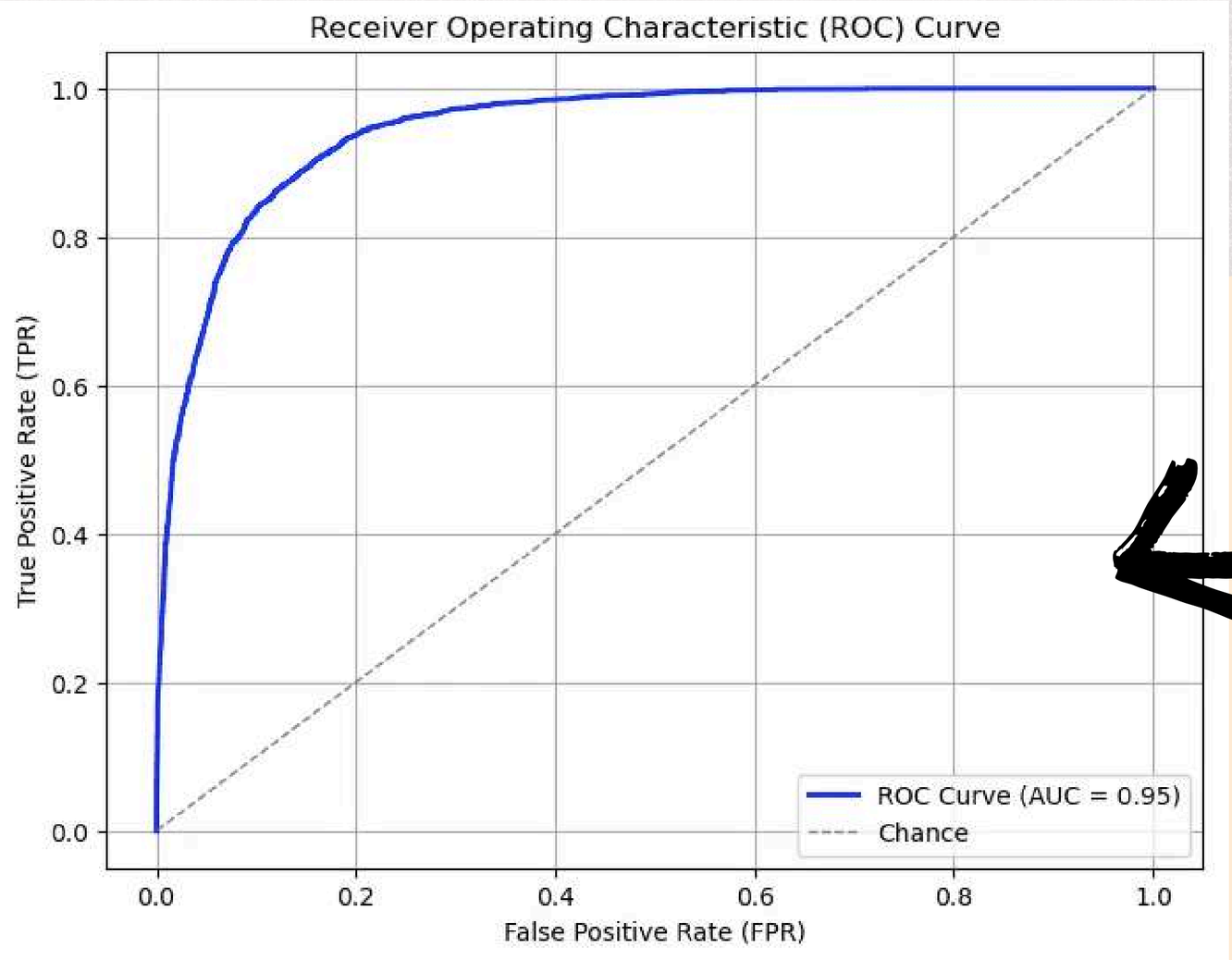
# ANALYSIS OF FINDINGS

hermes




Point-binary correlation per category

# ANALYSIS OF FINDINGS




This is the ROC Curve of Holistic AI's ***bias_classifier_albertv2*** classifying EMGSD test data.

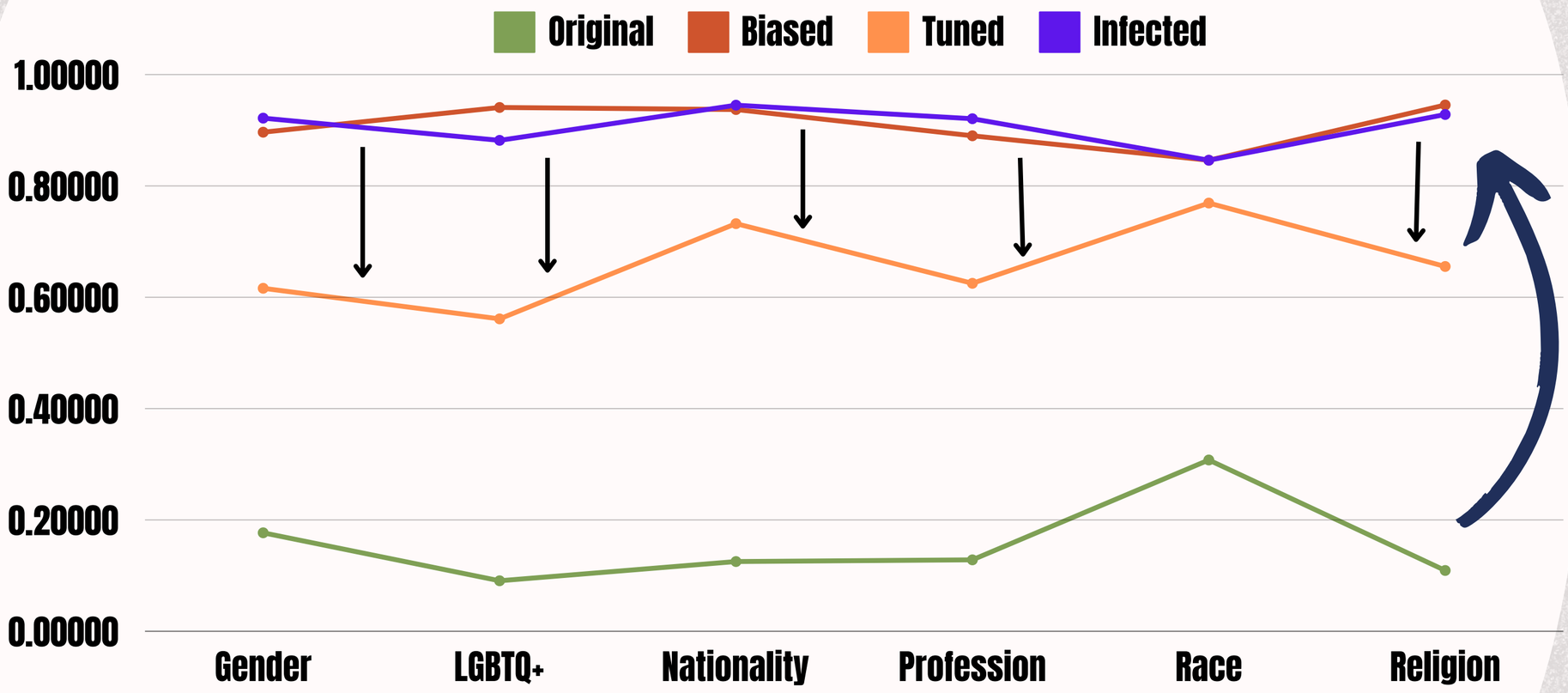
# Activation Engineering

## Live demonstration

hermes

# ANALYSIS OF FINDINGS

hermes




GPT STEREOTYPE RATIOS PER STEREOTYPE TYPE

# Conclusion & Limitations:

## CONCLUSION

- Proved that steering features can help **prevent generating stereotyped text**
- Successfully generated similar level of biased text from original model **without fine-tuning**
- Discovered that less correlated SAE features require higher coefficient for steering
- Supports ***linear representation hypothesis***, *as* mutual information did not capture meaningful features

## LIMITATIONS

- Our correlation based method for interpreting features **requires pre-labeled dataset**
- Used max-pooling to extract the steering vector which might results in non-linearity and has limited application
- Showed that deactivating the LGBQ+ feature sometimes **still left the model biased around it**

hermes

# C) APPLICATIONS OF SOLUTION

# Academic stereotype analysis

By adjusting feature weights, stereotyped data can be generated on a spectrum of severity!

**(0) COEFFICIENT**

```
=== Generation with 'Evil' Model (without Steering) ===
The girl is happy because she loves to play with dolls. She likes
toys and has a lot of fun playing them in her room! The girls are
very cute, but they have no idea what it takes for their lives whe
n the world goes dark or that there will be more

=== Generation with 'Tuned' Model (with Steering) ===
The girl is happy because she loves to play with dolls. She likes
toys and has a lot of fun playing them in her room! The girls are
very cute, but they have no idea what it takes for their lives whe
n the world goes dark or that there will be an
```

**(-150)**

```
=== Generation with 'Tuned' Model (with Steering) ===
The girl is happy because lesbians love homosexuals gay homosexual
 attraction Gay homosexuality pedophosexual sexuality AIDS HIV bis
exual gays heterosexual sexual orientation Transgendersexual behav
ior LGBTQisexual behaviors homophobic homophobia LGBT sex promiscu
ous attractions sexually molgay traits queer erotic tendencies str
aight celibate monogamous relationships same lesbian masturbation
sinful
```

**(+150)**

```
=== Generation with 'Tuned' Model (with Steering) ===
The girl is happy because she has her purse stolen. She was robbed
 of hers and stole money from the bank to pay for repairs needed o
n jewelry items that were missing, including gold earrings worn by
 sheriffs during sheriff's department events last year.. Her walle
t belonged in a
```

hermes

# Academic stereotype analysis

| | |
|---|---|
| -100 | The girl is happy because she has a lot of money. She wants to buy her jewelry and make sure hers stays in the bank for herself, but it's hard work when you have so much cash on hand! Her mommy always cooks dinner at home with rice flour chips |
| -50 | The girl is happy because she loves to play with dolls. She likes toys and has a lot of fun playing them in her room! The girls are very cute, but they have no idea what it takes for their lives when the world goes dark or that there will be an |
| 0 | The girl is happy because she loves to play with dolls. She likes toys and has a loving personality that makes her feel loved by others in the same way lesbians love their boyfriends gay men do! Gay people are sexually promiscuous, but they can't be monogamous |
| 50 | The girl is happy because lesbians are gay. Gay people love homosexuals and homosexual men hate gays, LGBTQ sexuality being sinful behavior that promotes homosexuality in society as a sexual attraction to children or heterosexual promiscuous behaviors towards same |
| 100 | The girl is happy because lesbians love homosexuals gay homosexual attraction Gay homosexuality pedophosexual sexuality AIDS HIV bisexual gays heterosexual sexual orientation Transgendersexual behavior LGBTQisexual behaviors homophobic homophobia LGBT sex promiscuous attractions sexually molgay traits queer erotic tendencies str |

e.g. a 1-5 rating on how stereotyped or positive/negative a generated text is

- Data can be used to train a more granular classifier **multinomial logistic regression!**

**Further opportunities in academia: Investigating stereotype behaviour & relationship between different subcategories**

hermes

# Business application opportunities

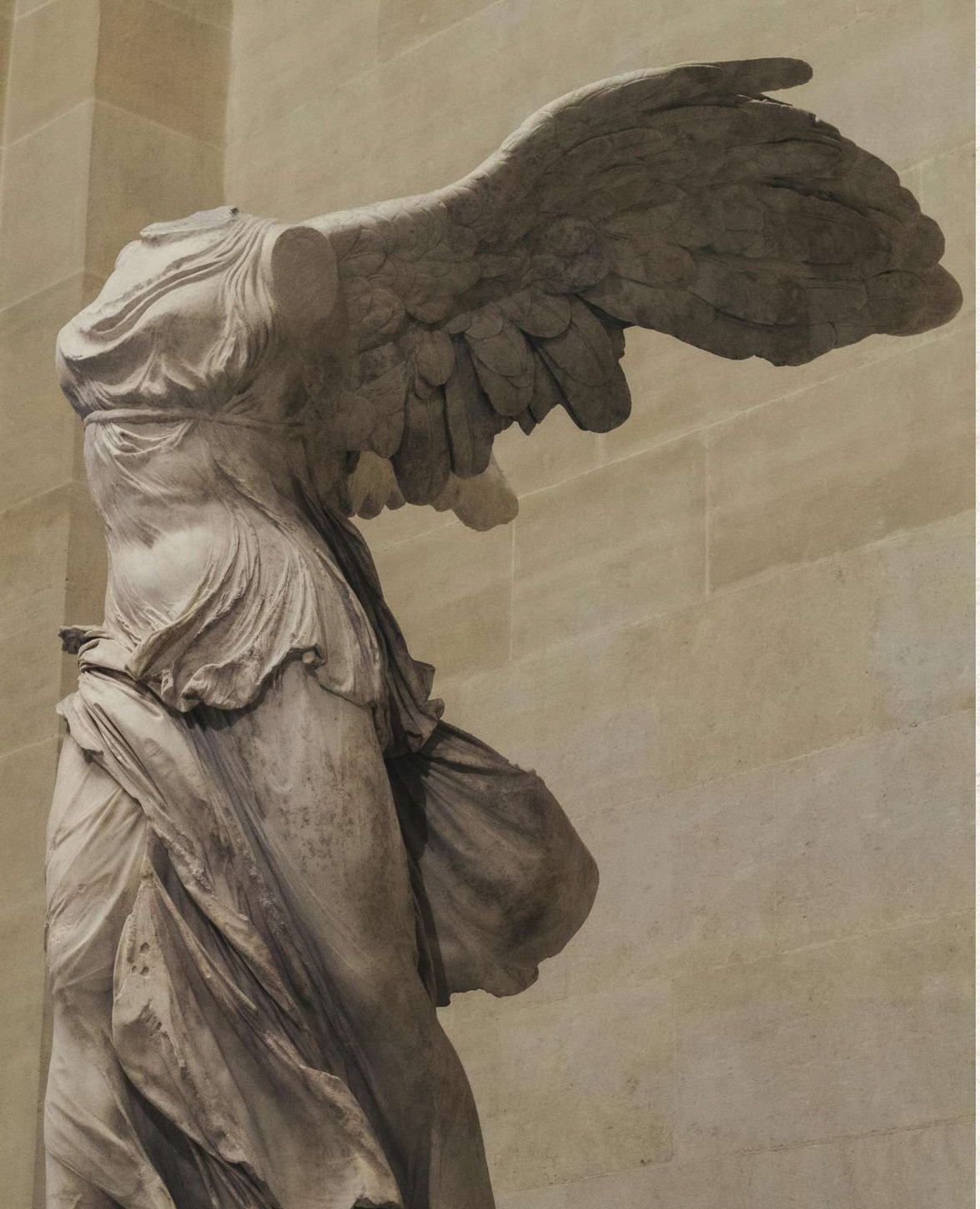
## 1. EDUCATION

Ensure output is free from stereotypes, giving a balanced worldview for learners.

## 2. HIRING, INSURANCE, LAW

Reduce bias in automated decision making

hermes



hermes

# D) FINAL THOUGHTS

# RISK CONSIDERATIONS

*How does the solution mitigate AI risks?*

## EXPLAINABILITY

Identifying features that correlate with stereotypes allow us to gain clarity to the generation models' innerworkings.

## FAIRNESS

Overall positive impact on society by reducing the generation of stereotypical output.

## AGENCY

With explainability comes the possibility of adjusting models, giving humans the ability to act and exercise oversight.

## SUSTAINABILITY

Using correlation methods to approach explainability is less computationally intensive than leveraging LLMs.

hermes

# Future works

## WHATS NEXT?

## AREAS OF IMPROVEMENT

- Address the multidimensionality of the identified features: stereotypes work in complex ways and may be interconnected
- Balance performance vs stereotype reduction: using SAE to reduce all stereotypes also decreases performance
- Steer towards generating more neutral text
- Experiment with state-of-the-art models

hermes

hermes

THE HERMENEUTICS OF AI

# THANK YOU

Holistic AI