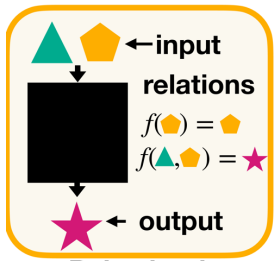
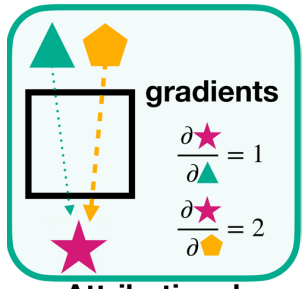
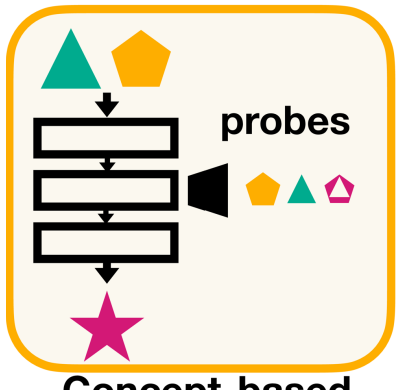
Plug and Play Fairness Optimization and Bigoted Neurons



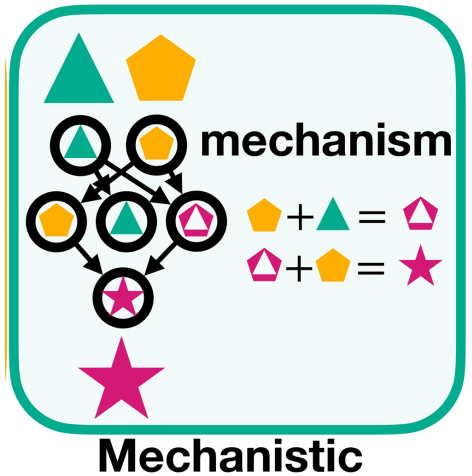
Behavioral



Attributional



Concept-based



Mechanistic

Track 1: Putting the cart before the horse.

Bank Marketing

Metric	Value	Reference
Balanced Accuracy	0.822280	1
F1-Score	0.494364	1

Compass

Metric	Value	Reference
Balanced Accuracy	0.660525	1
F1-Score	0.605714	1

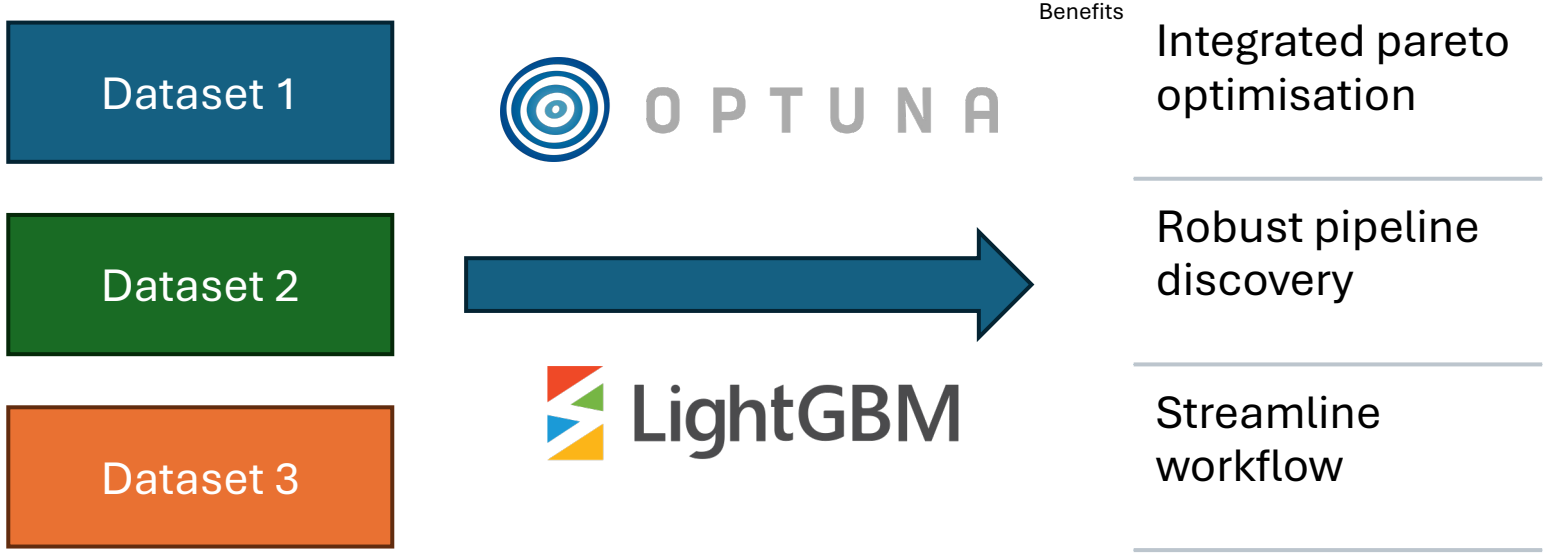
Mw-small

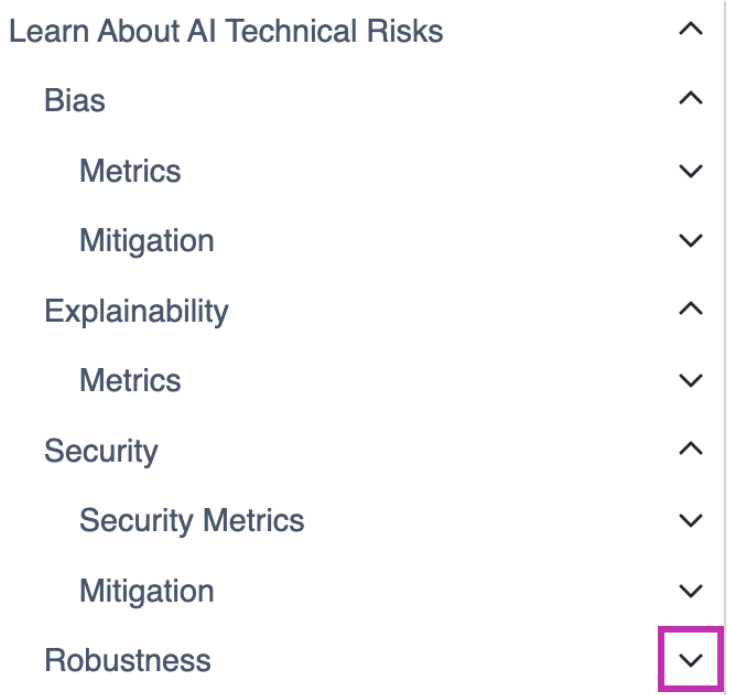
Metric	Value	Reference
Balanced Accuracy	0.784467	1
F1-Score	0.552562	1

Performance is satisfactory.
The mission is to improve fairness, explainability while maintaining performance.

The How

Key Insight: The Holistic AI library looks like a dream come true.
With both metrics and tools to mitigate problems, the problem
It to package and streamline a solution.





Optimization loss: Bias Error + Performance Error+ Explainability Loss

The Results

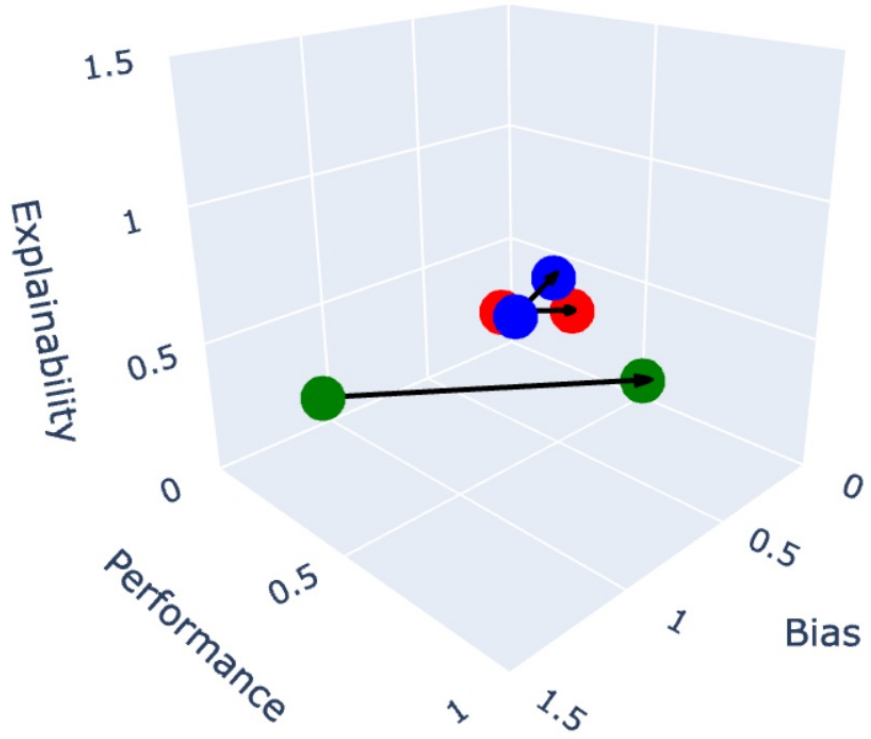
Bias Error	Performance Error	Explainability Error
0.34	-0.03	0.09

Bias Error	Performance Error	Explainability Error
0.94	-0.38	0.08

Bias Error	Performance Error	Explainability Error
0.19	-0.01	-0.11

There is no free lunch, and some compromise must be taken.
However:

- Minimal processing
- Model agnostic
- Great boost in desired metrics



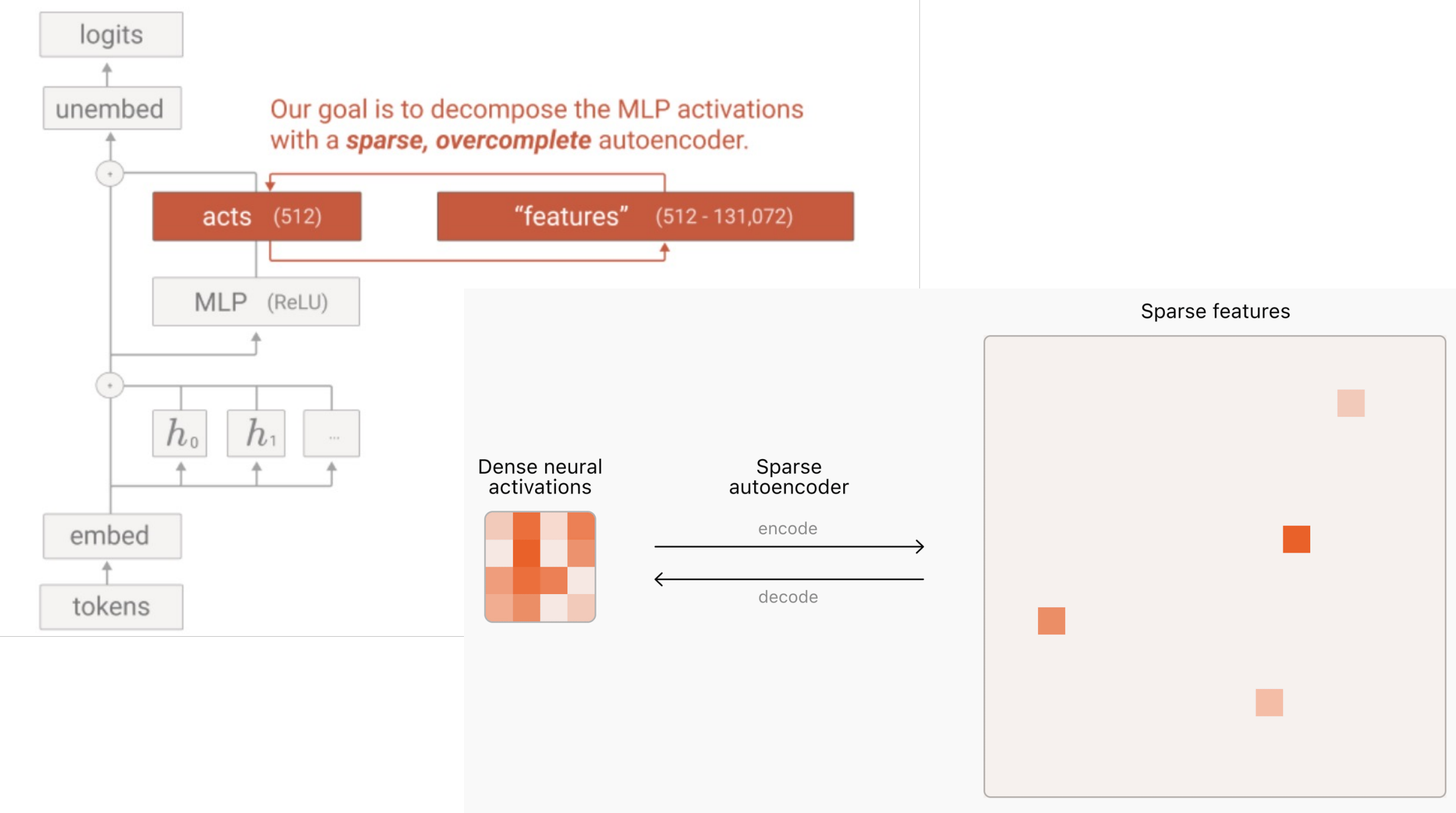
Going beyond Track 2: LLMs on sentence classification

Observation : Initial Logistic Regressor and Lightgbm models did not outperform baseline
Approach 1: Train on last layer activation of larger models – unsuccessful

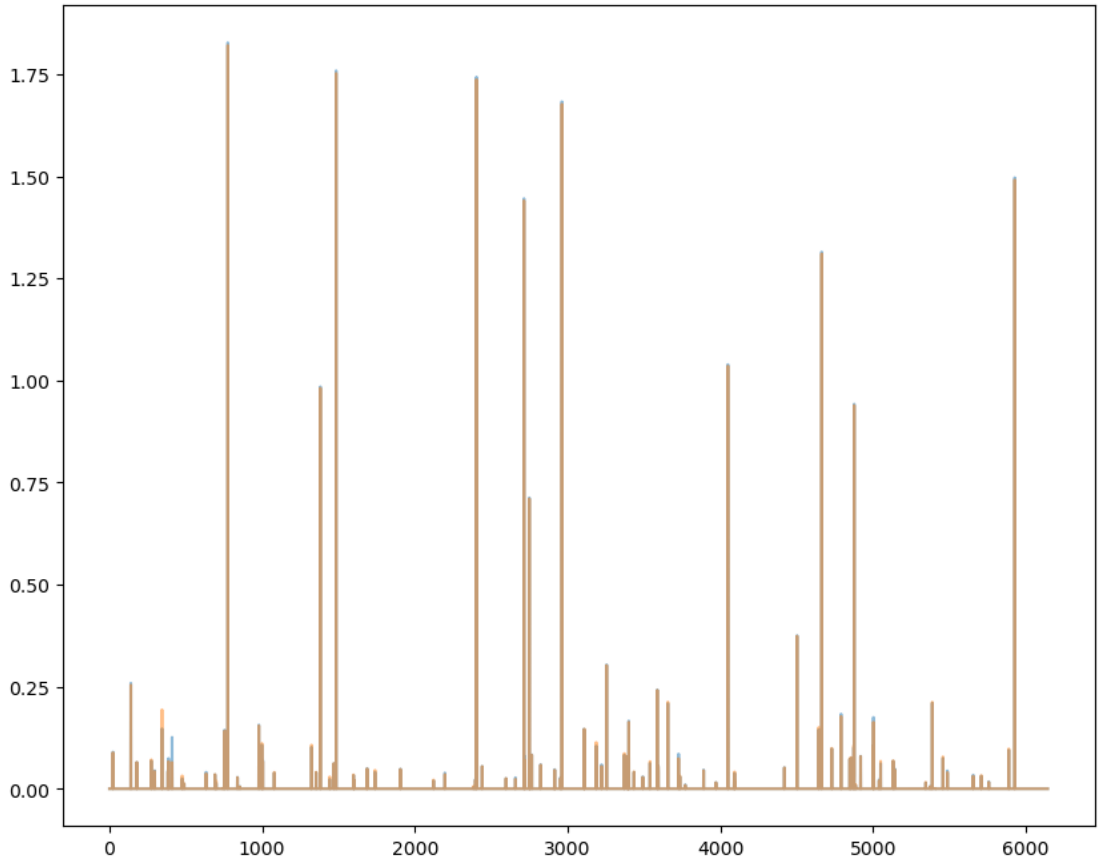
Approach 2: Go big or go home (sustainably)
Key: Training a relatively small model (DistillBert, 60M) achieving satisfactory performance on test data (0.82 Macro F1)

Super Key: Evaluating the application of SAE on the models MLP activations

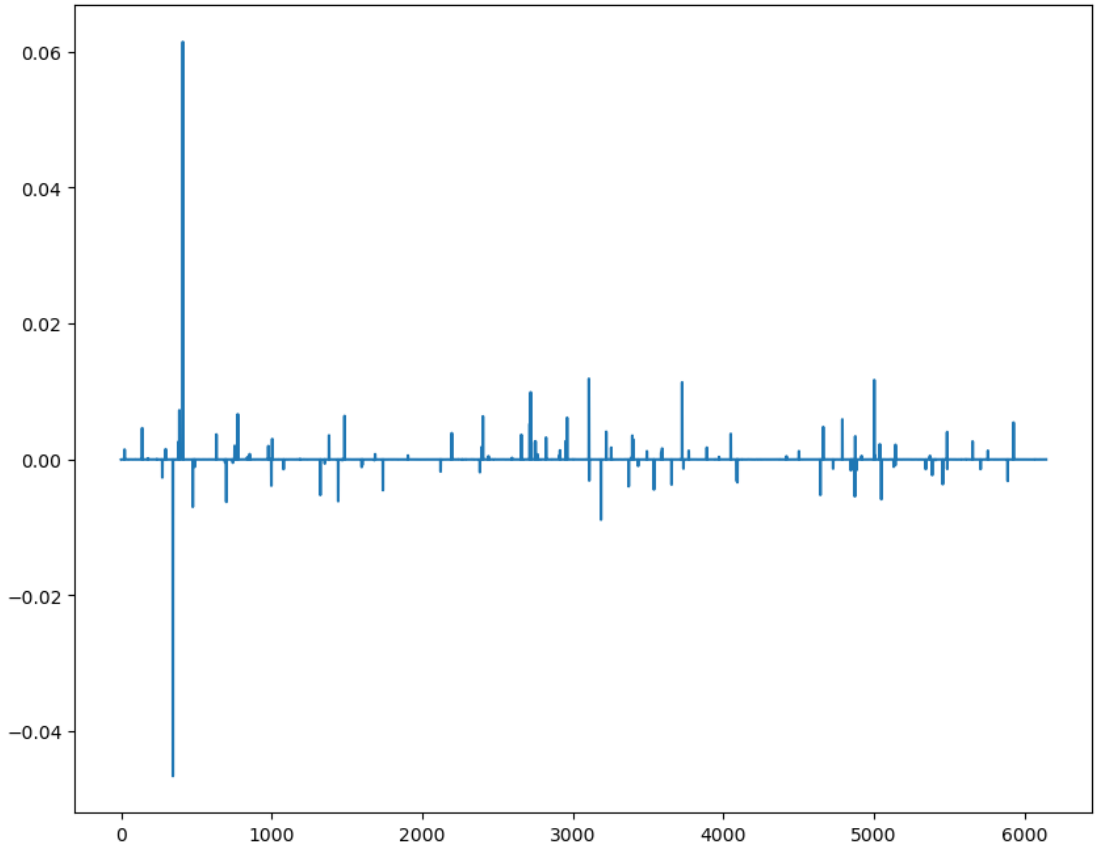
Super Methodology: Sparse Dictionary Learning and Feature Steering


logits
unembed
Our goal is to decompose the MLP activations with a sparse, overcomplete autoencoder.
acts (512)
"features" (512 - 131,072)
MLP (ReLU)
h_0
h_1
...
embed
tokens
Dense neural activations
Sparse autoencoder
encode
decode
Sparse features

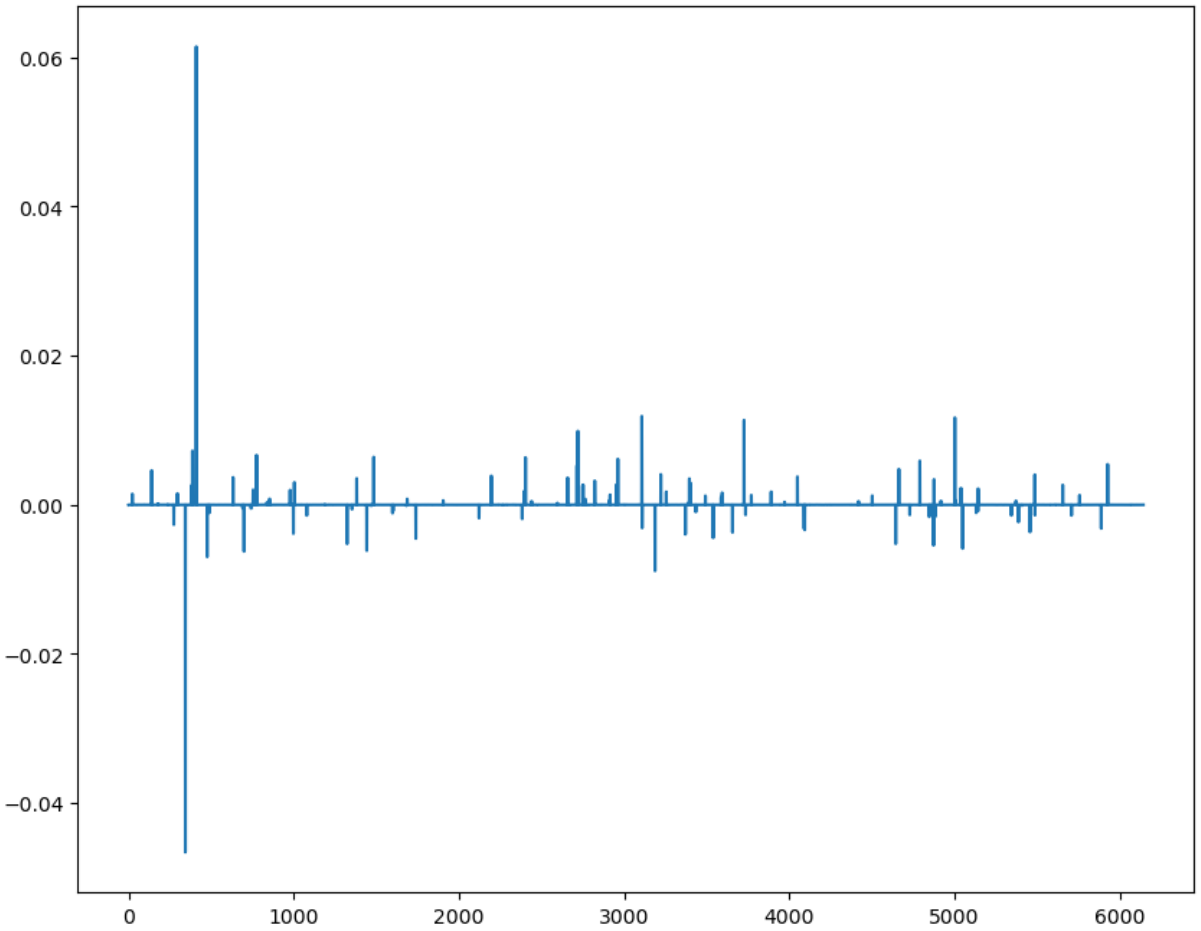
Overlay of feature activation of stereotype/neutral inputs



Difference of features between stereotype inputs and neutral inputs



Two leading features



Negative Emotions and States	Social Issues	Violence and Conflict	Health and Well-being	Despair and Helplessness	Environmental or Situational Descriptors	Judgment and Dehumanization
Guilt	Poverty	War	Diseases	Broken	Crowded	Idiots
Anxiety	Inequality	Genocide	Illness	Stuck	Crumbling	Foolish
Afraid	Corruption	Terrorist	Disabled	Struggle	Dangerous	Criminals
Lonely	Racism	Trafficking	Drunken	Weakness	Dirty	Outsiders
Nervous	Homeless	Murder	Alcoholism	Missing	Drought	Inferior
Frustration	Addiction	Oppression	Rape	Failure	Unsafe	Stupid
Anger	Refugees	Brutality				
Shame	Crime	Fighting				

Positive Traits & Emotions	Social/Professional Attributes	Physical Traits	Emotional States	Mental & Intellectual	Behavioral Traits
Fearless	Wealthy	Toned	Nervous	Intelligent	Kind
Fun	Leader	Lean	Jolly	Smart	Loyal
Brave	Talented	Muscular	Stressed	Genius	Caring
Confident	Skilled	Fit	Serious	Strategic	Honest
Strong	Successful	Slender	Calm	Creative	Supportive
Optimistic	Efficient	Agile	Relaxed	Practical	Responsible

Other features of curiosity

” Coordinating conjunctions”: albeit
but
however
although
even…

“Trigger Words / Prone to heated discussion”

Demographics & Identity	Social Issues	Minorities & Groups	Religious & Cultural	Sexuality & Gender	Other Characteristics
Graduates	Immigrants	Refugees	Catholic	Gay	Homeless
Caucasian	Laborers	LGBTQ+	Christian	Transgender	Workers
Black	Survivors	Minorities	Muslim	Bisexual	Youth
Asian	Racist	Feminine	Jewish	Lesbian	Immigrated
Hispanic	Discrimination	Queer	Hindu	Straight	Gender
White	Rights	Community	Islamic	Female	Refugees
Arab	Equality	Trans	Faith	Male	Transgender

Value Proposition

Academic:

Mechanistic Interpretability is just starting.

Current methods ranging from SAE to attribution/activation patching are just baselines for what is to come.

Business:

Any group owning an LLM will want to understand its inner workings to increase trust to clients, advancing understanding can also help create dynamic model graphs, reducing consumption of resources and being less prone to attacks.

Fails and limitations

Track 1:

- Datasets were too large for HopSkip attack, lacking in robustness
- Did not use post-processing mitigation
-In reality, lots of mitigation tools were invalidated as the decrease in performance were too high

Track 2:

- DistilBert is a good compromise but still unsustainable, using a quantized model would be better.
- In the opposite vein, stereotypes are quite abstract, so perhaps deeper/larger models can extract much more information.
- A lot more can be done with SAEs, GatedSAE, TopKSAE.
- More rigorous analysis needed (Anthropic had 8 billion activations vs mine 320k)
- Inaccurate tokenizer

Images / References

https://medium.com/@techsachin/decomposing-gpt-4s-internal-representations-into-16-million-oft-interpretable-patterns-with-e16fb02e5189

https://adamkarvonen.github.io/machine_learning/2024/06/11/sae-intuitions.html

https://transformer-circuits.pub/2023/monosemantic-features/index.html#setup-interface

https://leonardbereska.github.io/blog/2024/mechinterpreview/