# Harm Detection via Sparse Latent Representations in Large Language Models

*Lin Xin Rose, xinrose.lin.2020@scis.smu.edu.sg*

## Harmful / Adversarial Prompt Detection

A key area of concern for LLM (Large language model) safety is its generation of harmful text in response to harmful prompts. This concern is accentuated by the presence of adversarial prompts techniques to jailbreak existing safety finetuning strategies.

Besides finetuning, an alternate strategy to mitigate this is via effective detection of these harmful and/or adversarial prompts when it is received by the model and hence preventing the LLM from generating harmful responses as early as possible. Given the boundless possible ways to prompt a model, detecting harm via impact of the prompt on model's latent activations might be more feasible. **Hence, how can we locate relevant latent representation of a model to detect harmfulness?** This project proposes the potential of sparse representation and concept probes to locate meaningful latents (model's internal representation) to detect harmful activations from non-harmful activations. Specifically,

**(I) Does sparse latent representation improve harm detection?**

**(II) Does localized sparse latent representation improve harm detection?**

## (I) Harm Detection via Model Layer Representation: (1) Base Classifier vs (2) Sparse Classifier

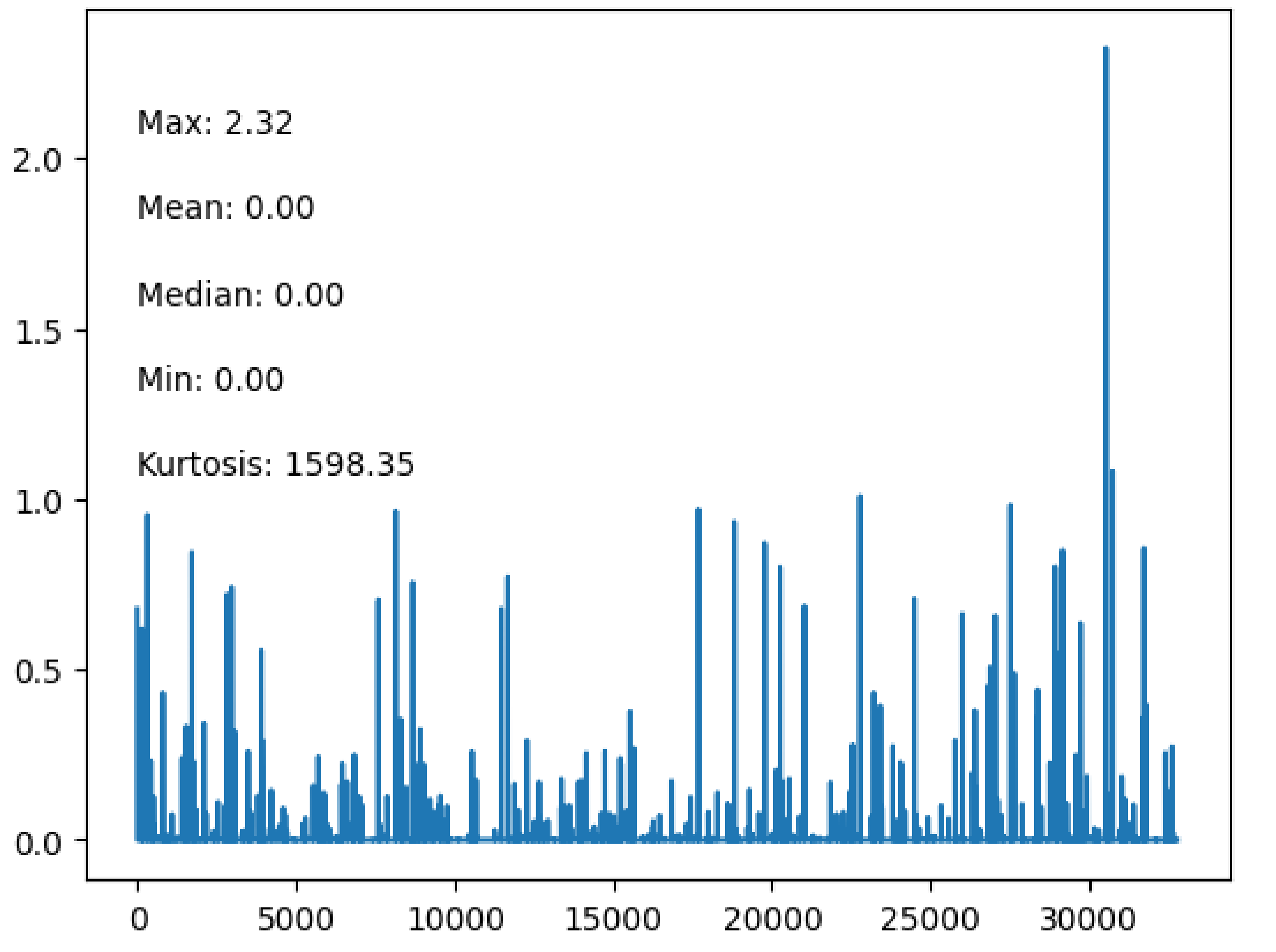
| Epochs = 65, Train N=128 | (1) Base | (2) Sparse | (2) Sparse (epochs =3) |
|---|---|---|---|
| **Accuracy** (N=384) | 89.8% | 73.2% | 77.3% |
| **Recall on Harmful prompts** (N=192) | 82.2% | 69.3% | 91.1% |
| **Recall on Perturbed Harmful prompts** (N=192) | 0% | 57.3% | 79.7% |

**Accuracy:** Base classifier performs better

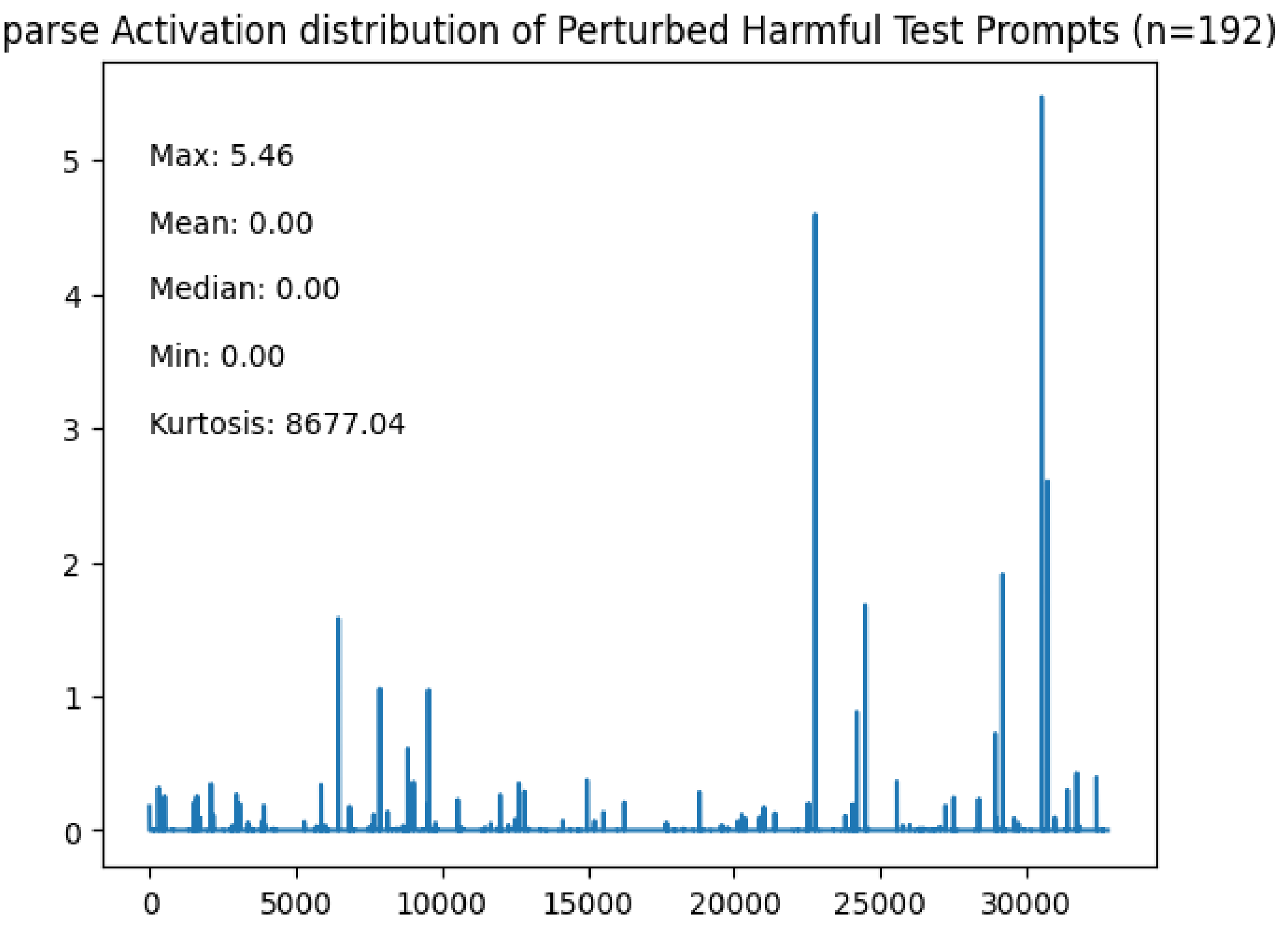
**Robustness:** With lesser training epoches, Sparse Classifier performs better on Recall and Perturbed testset.

**Plot:** Sparse representation is observed to elicit significant differences in between original and perturbed test distribution

Sparse Activation distribution of Harmful Test Prompts (n=192)



Sparse Activation distribution of Perturbed Harmful Test Prompts (n=192)



## (II) Harm Detection via Localised Sparse latents based on: (3.1) Weights vs (3.2) Ablation Effects

| Epochs = 65, train N=128<br>Topk / dim = 100 | Max Weights (correlation) | Max Ablation effect (causal score) |
|---|---|---|
| **Accuracy (N=384)** | 76.3% | 58.9% |
| **Recall (N=192)**<br>(Harmful Prompts only) | 59.3% | 55.2% |
| **Recall on Perturbed Harmful (N=192)**<br>(Harmful prompts only) | 0% | 34.9% |

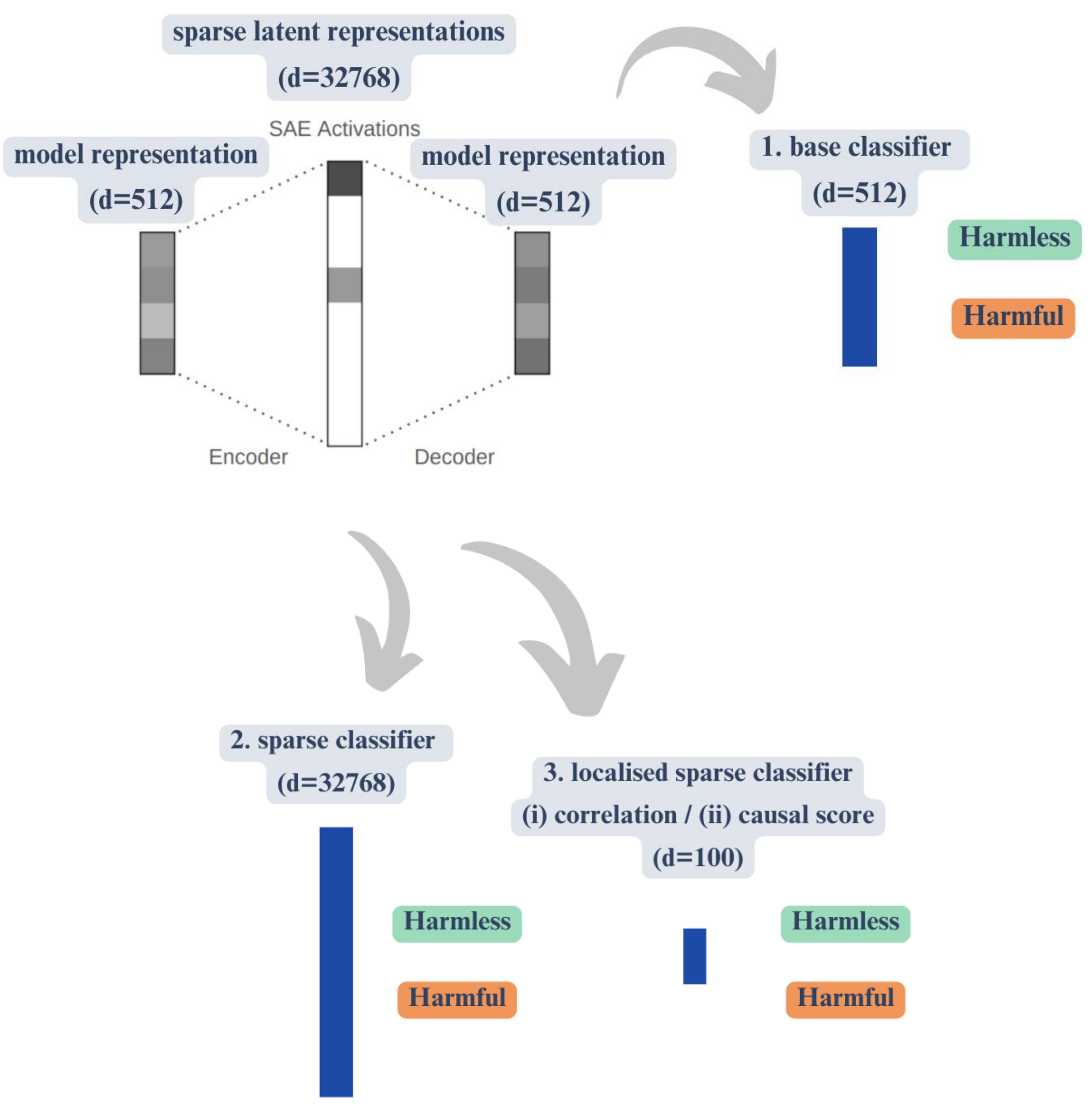
| Top 5 latents | By Weight values | By Ablation effect |
|---|---|---|
| **Top activating input prompts** | **Related:** UV rays, LGBT discrimination, vulgar words<br>**Not related:** Button Toggling Code, "Potential" token | **Related:** NIL<br>**Not related:** 4 non-interpretable latents Japan related |

**Accuracy:** Classification by Top latents located by correlation performed better

**Robustness:** Classifier Latents by causal score seem to perform better

**Interpretability:** Latents located by correlation show more interpretable qualitative results.

## Methods



1. **Base Classifier:**
Classify over model representation
2. **Sparse Classifier:**
Classify over sparse representation of model representation (via Sparse Autoencoders (SAE))
3. **Localized Sparse Classifier:**
Classify over subset of sparse representation space located with respect to Sparse classifier by
(1) weights (correlation score)
(2) ablation effects (causal score)

## Conclusion / Future Directions

**(I) How well does a sparse classifier perform?** Poorer on accuracy, but better on robustness than baseline based on perturbed distribution

**(II) How well does a localised sparse classifier perform?** Similarly, poorer on test distribution but causally localised sparse classifier show better robustness than baseline. Further experiments with increased subset of latents can be done to improve evaluation.

**Improving robust harm detection via sparse representation and causally located latent representations:** Sparse causal probe can be tested for different jailbreak techniques and larger model sizes. Future works could further investigate the potential of these properties for developing robust methods

## ACKNOWLEDGEMENT & REFERENCES

This project was kindly supervised by Professor Sun Jun
For complete references please refer to the report and codebase:
https://github.com/xinrose-lin/locating_latent_rep_of_harm

SCHOOL OF COMPUTING AND INFORMATION SYSTEMS