

# Language-Independent Concept Spaces in Large Language Models

*Under advice of Greg Anderson*

Aidan M. Mokalla

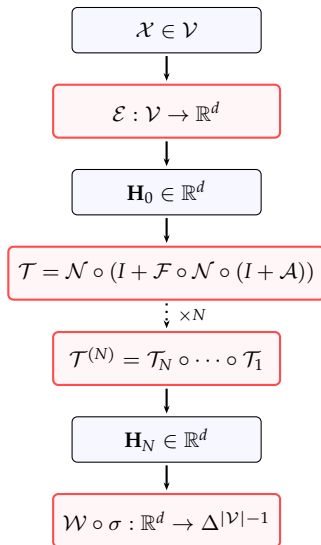
Reed College

AidanM@reed.edu

Oral Thesis Exam  
July 14, 2025

# Interpreting Large Language Models

# Interpreting Large Language Models



# Interpreting Large Language Models

## 1. Safety & Alignment<sup>a</sup>

*<sup>a</sup>"We check whether Sparse Autoencoders find features correlated with lying and deception in this out-of-distribution setting."*  
doi.org/10.48550/arXiv.2504.04072

# Interpreting Large Language Models

## 1. Safety & Alignment

## 2. Bias & Fairness<sup>a</sup>

---

*<sup>a</sup>"we receive mixed signals as only some subsets of the data are useful in providing insights. To alleviate these two problems, we introduce a more rigorous evaluation dataset and a debiasing method based on Sparse Autoencoders to help reduce bias in models."*

[doi.org/10.48550/arXiv.2410.13146](https://doi.org/10.48550/arXiv.2410.13146)

# Interpreting Large Language Models

1. Safety & Alignment
2. Bias & Fairness
3. Compliance with Policies<sup>a</sup>

---

<sup>a</sup>*"These goals—the targeted removal of information from a model and the targeted suppression of information from a model's outputs—present various technical and substantive challenges."*  
[doi.org/10.48550/arXiv.2412.06966](https://doi.org/10.48550/arXiv.2412.06966)

# Interpreting Large Language Models

1. Safety & Alignment
2. Bias & Fairness
3. Policy & compliance
4. Improving Performance<sup>a</sup>

---

<sup>a</sup>*"We also find more abstract features—responding to things like bugs in computer code."*  
[anthropic.com/research/mapping-mind-language-model](https://anthropic.com/research/mapping-mind-language-model)

# Interpreting Large Language Models

1. Safety & Alignment
2. Bias & Fairness
3. Policy & compliance
4. Improving Performance
5. Scientific Insight<sup>a</sup>

---

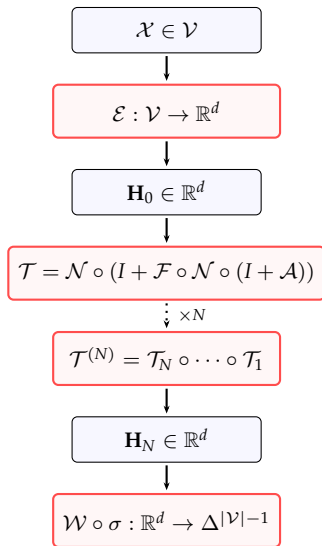
<sup>a</sup>*"We introduce a non-invasive decoder that reconstructs continuous language from cortical semantic representations recorded using functional magnetic resonance imaging (fMRI). Given novel brain recordings, this decoder generates intelligible word sequences that recover the meaning of perceived speech, imagined speech and even silent videos."*  
[doi.org/10.1038/s41593-023-01304-9](https://doi.org/10.1038/s41593-023-01304-9)



# Interpreting Large Language Models

Idea: *Concept Space*

$$\mathcal{C} \stackrel{\text{def}}{=} \text{span} \left[ \bigcup_{\text{concepts}} \right]$$



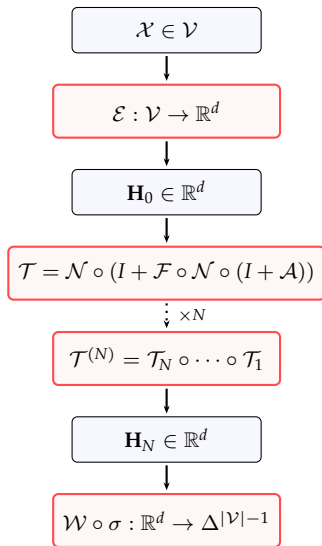
# Interpreting Large Language Models

Idea: *Concept Space*

$$\mathcal{C} \stackrel{\text{def}}{=} \text{span} \left[ \bigcup_{\text{concepts}} \right]$$

**Where is  $\mathcal{C}$ ?**

1.  $\mathbb{R}^d \stackrel{?}{\cong} \mathcal{C}$



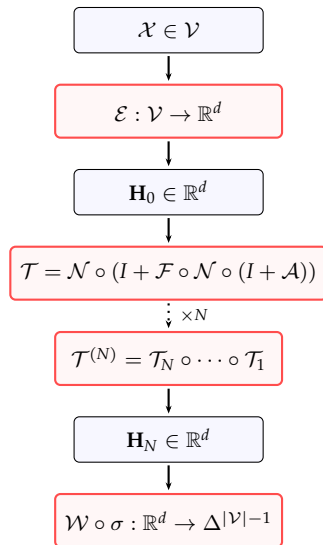
# Interpreting Large Language Models

Idea: *Concept Space*

$$\mathcal{C} \stackrel{\text{def}}{=} \text{span} \left[ \bigcup_{\text{concepts}} \right]$$

**Where is  $\mathcal{C}$ ?**

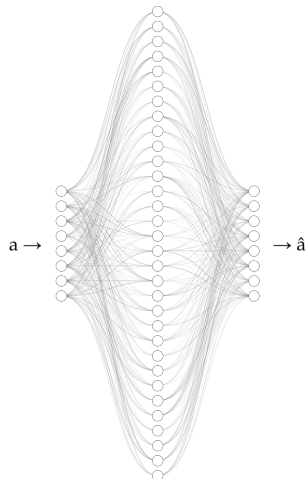
1.  $\mathbb{R}^d \not\cong \mathcal{C}$
2.  $\mathbb{R}^{d^{x>1}} \cong \mathcal{C}$



# Approximating $\mathcal{C}$ with Sparse Coding

SAE:  $\mathbb{R}^d \rightarrow \mathbb{R}^{d^{>1}} \rightarrow \mathbb{R}^d$

## 1. Sparse Autoencoders (SAEs)

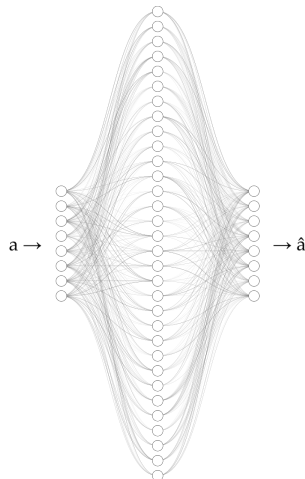


# Approximating $\mathcal{C}$ with Sparse Coding

SAE:  $\mathbb{R}^d \rightarrow \mathbb{R}^{d^{>1}} \rightarrow \mathbb{R}^d$

## 1. Sparse Autoencoders (SAEs)

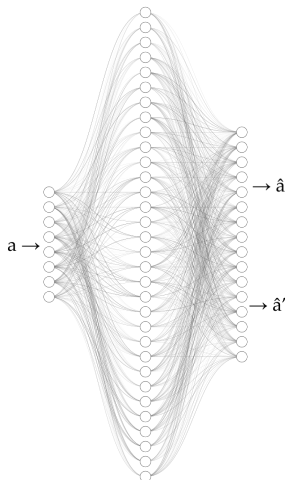
$$\begin{aligned} \mathcal{L}(\text{SAE}) = & \\ & \|a - \hat{a}\|_2^2 \\ & + \lambda \text{sparsity}(\text{SAE}, a) \end{aligned}$$



# Approximating $\mathcal{C}$ with Sparse Coding

$$\text{USAE: } \mathbb{R}^d \rightarrow \mathbb{R}^{d^{x>1}} \rightarrow \mathbb{R}^{2d}$$

1. **Sparse Autoencoders**
2. **Universal Sparse Autoencoders (USAEs)**

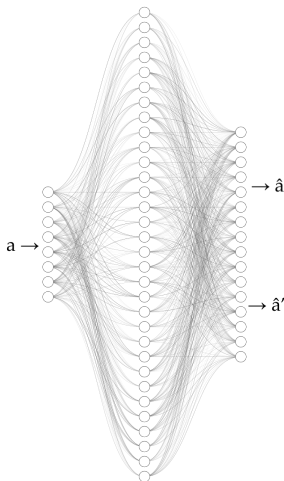


# Approximating $\mathcal{C}$ with Sparse Coding

$$\text{USAE: } \mathbb{R}^d \rightarrow \mathbb{R}^{d^{x>1}} \rightarrow \mathbb{R}^{2d}$$

1. **Sparse Autoencoders**
2. **Universal Sparse Autoencoders (USAEs)**

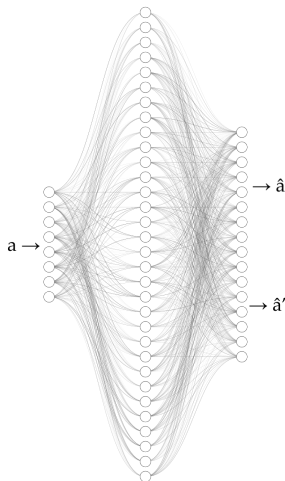
$$\begin{aligned} \mathcal{L}(\text{USAE}) = & \\ & \|a - \hat{a}\|_2^2 \\ & + \lambda \text{sparsity}(\text{USAE}, a) \\ & + \lambda' \|a' - \hat{a}'\|_2^2 \end{aligned}$$



# Approximating $\mathcal{C}$ with Sparse Coding

$$\text{USAE: } \mathbb{R}^d \rightarrow \mathbb{R}^{d^{x>1}} \rightarrow \mathbb{R}^{2d}$$

1.  $\mathcal{L}(\text{SAE}) =$   
 $\|a - \hat{a}\|_2^2$   
 $+ \lambda \text{sparsity}(\text{SAE}, a)$
2.  $\mathcal{L}(\text{USAE}) =$   
 $\|a - \hat{a}\|_2^2$   
 $+ \lambda \text{sparsity}(\text{USAE}, a)$   
 $+ \lambda' \|a' - \hat{a}'\|_2^2$
3. USAEs are more interpretable than SAEs.





# Approximating $\mathcal{C}$ with Sparse Coding

3. **USAEs** are more interpretable than **SAEs**.

[i] **Concept  
Convergence**

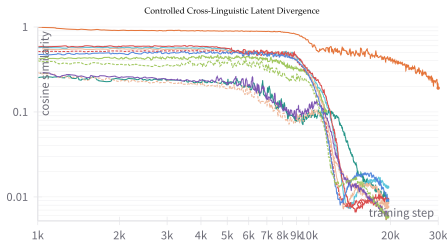
[ii] **Reconstruction  
Accuracy**

[iii] **Sparsity**

# Approximating $\mathcal{C}$ with Sparse Coding

3. **USAEs** are more interpretable than **SAEs**.

## [i] Concept convergence



## [ii] Reconstruction

Accuracy

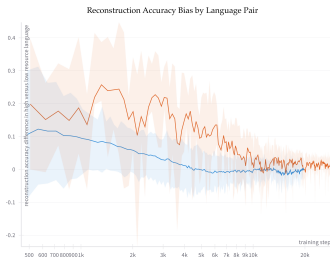
## [iii] Sparsity

# Approximating $\mathcal{C}$ with Sparse Coding

3. **USAEs** are more interpretable than **SAEs**.

[i] Concept convergence

[ii] Reconstruction Accuracy

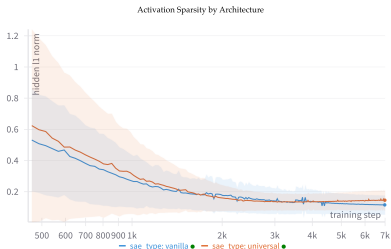


[iii] Sparsity

# Approximating $\mathcal{C}$ with Sparse Coding

3. **USAEs** are more interpretable than **SAEs**.

- [i] Concept convergence
- [ii] Reconstruction Accuracy
- [iii] Sparsity



Fin