INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026 / Main Track

Diffusion-Time Concept Manifolds

Sparse Autoencoder Groups for Interpreting Denoising Language Models

Pawan Kumar IIIT, Hyderabad, 500032

Can small feature groups reveal sentiment as masking changes?

aiwranglers.org optimalstar.org

- We study how language models represent sentiment, from negative to positive.
- We look for small groups of sparse autoencoder features that carry this information.
- We test how well these groups predict sentiment, how stable they are, and what happens when we edit them.

What changes when words are hidden?

MASKED LANGUAGE DIFFUSION

The model uses context to recover hidden words.

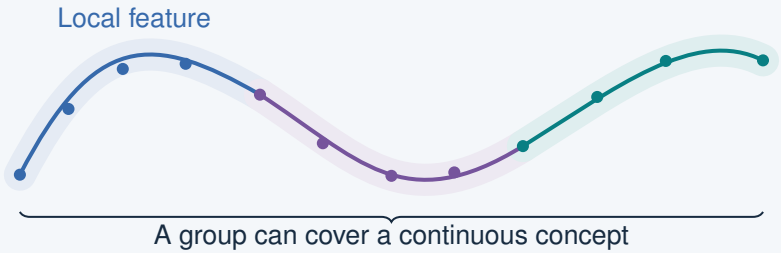
Masked	[MASK]	[MASK]	was	[MASK]
Partial	The	film	was	[MASK]
Complete	The	film	was	wonderful

Illustration only. Our static tests use independently masked inputs.

WHY USE A GROUP?

One feature may capture only part of the sentiment range.

Several features can work together.



Sources: Sahoo et al., 2024; Nie et al., 2025; Gao et al., 2024; Bhalla et al., 2026; Wang et al., 2026. Full references in backup slides.

- Masking hides words. The model uses the remaining context to recover them.
- We test inputs at four masking levels: 15, 35, 60, and 85 percent.
- Different features can become useful as masking changes, so we compare groups across these levels.

How we form a sentiment feature group

01 / LEARN

Train the SAE to reconstruct hidden states with a few active features.

No sentiment labels.

→

02 / RANK

Rank features by how strongly they track training sentiment scores.

Positive or negative correlation.

→

03 / GROUP

Select top-ranked feature IDs.

Groups of 8, 16, 32, or 64.

Test each size.

→

04 / TEST

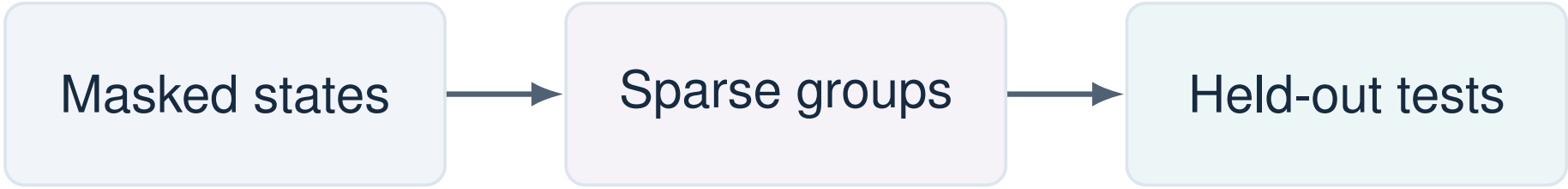
Fit a sentiment predictor on the selected activations.

Test on a reserved sentence pattern.

We also try groups based on features activating together or having statistical relationships: co-activation, graphical lasso, and Ising-style groups.

- The SAE learns features without sentiment labels. Only a few features activate for each input.
- We rank the features using training sentiment scores, then select groups of 8, 16, 32, or 64.
- We train a predictor for each group and test it on a reserved sentence pattern. Group members need not all activate together.

Four questions about each feature group



01 / PREDICTION

Can it predict sentiment on a new sentence pattern?

02 / GEOMETRY

Do similar sentiments have similar feature patterns?

03 / STABILITY

Do similar feature IDs remain relevant as masking changes?

04 / EDITING

Does editing the group change the model's output?

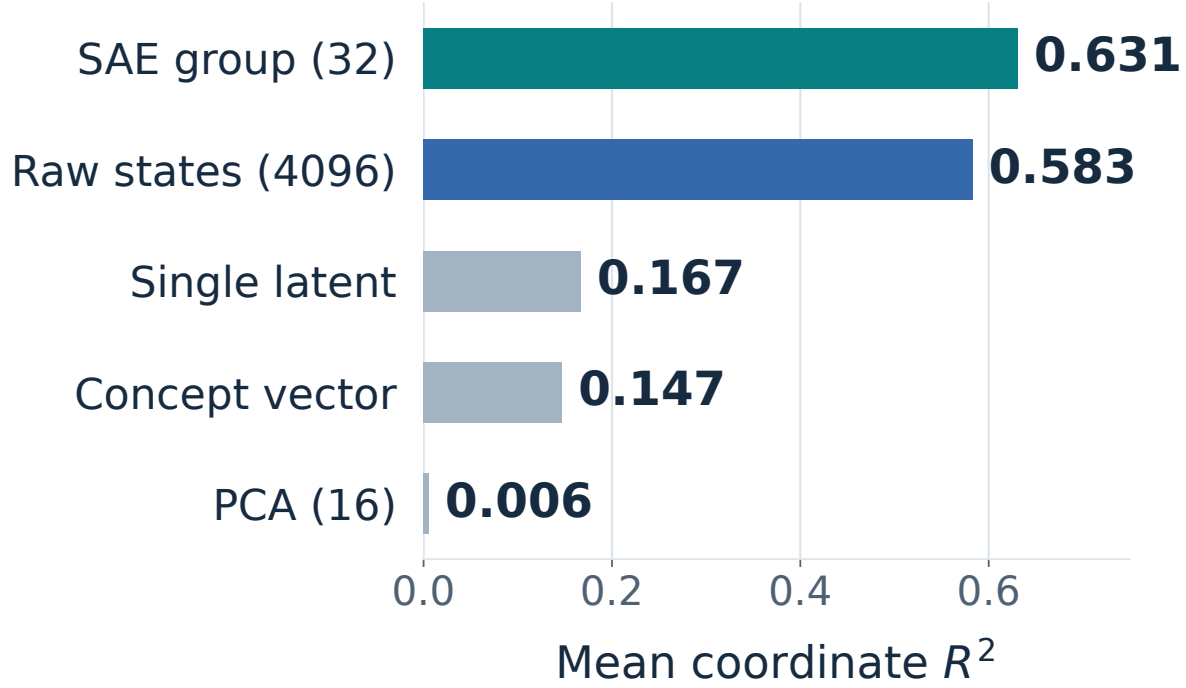
Group selection and predictor training use only the training patterns.

The SAE and input scaling stay fixed from earlier training; they may have seen the reserved pattern.

- We check whether groups predict sentiment and whether similar sentiments have similar feature patterns.
- We also compare groups across masking levels and test what changes when we edit them.
- These tests answer different questions. Good prediction alone does not show that the model needs those features.

32 features recover sentiment information

LLaDA layer 8; 15% masking; 3 seeds × 5 reserved sentence patterns



ABOUT 63% OF SCORE VARIATION

$R^2 = 0.631$

What each predictor uses

Raw state: all 4,096 dimensions.

Single latent: one SAE activation.

Concept vector: one projection.

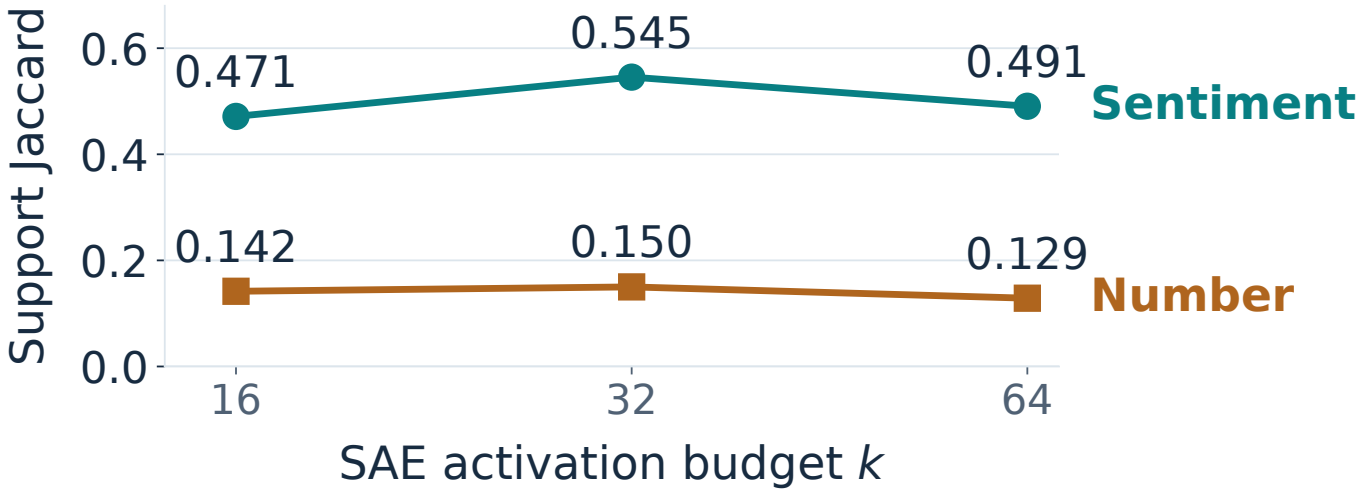
SAE group: 32 activations.

- Across 15 evaluations, the 32-feature group explains about 63 percent of the variation in sentiment scores.
- Raw states explain about 58 percent, while a single selected SAE feature explains about 17 percent.
- This is our strongest setting. Results vary across layers and masking levels; the full-grid mean R-squared is 0.212.

Stable feature groups; mixed results on new data

ACROSS MASKING LEVELS

Sentiment groups share more feature IDs.



Layer 8, seed 31; means over six mask pairs, not intervals across seeds.

NEW SENTENCES

Paraphrases: predict a score

Sparse / raw R^2: **.196** / **.087**

BERT layer 3; 840 sentences, 10 new templates.

Reviews: positive or negative

Sparse / raw AUC: **.598** / **.698**

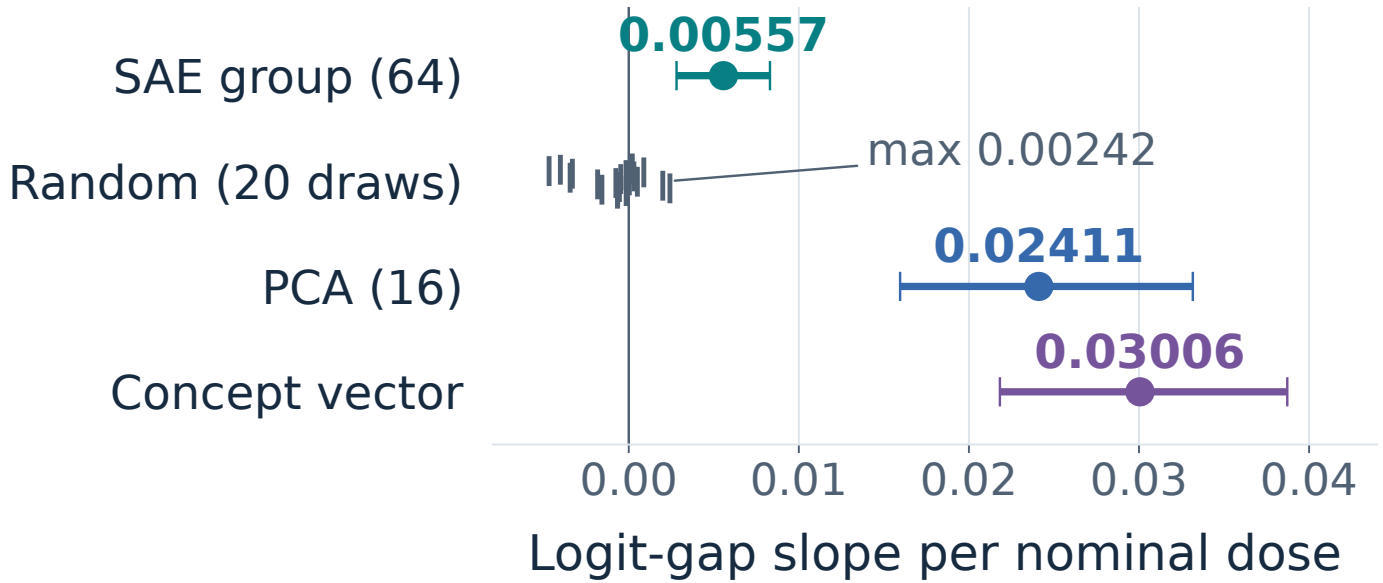
LLaDA layer 8, mask .15; trained and tested within SST-2. This is not zero-shot transfer.

- Many sentiment-related feature IDs remain relevant across masking levels. Prediction quality can still change.
- On new paraphrases, sparse features predict sentiment scores better than raw states.
- On natural reviews, raw states predict positive versus negative sentiment better. That test trains within the review dataset.

From sentiment prediction to targeted interventions

BERT layer 3; 64 selected features; 20 unseen fill-in-the-blank prompts

Farther right = stronger shift toward positive words.



Bars show 95% prompt-bootstrap confidence intervals. Equal nominal doses do not give equal changes to the actual hidden state.

- Editing the sentiment-related features nudges the model toward positive words. This response is stronger than for any of the twenty random feature groups we tested.
- The slope measures how much the model's preference for positive words changes as edit strength increases.
- PCA and concept-vector edits give larger shifts, but the actual hidden-state edit sizes are not equal.

TAKEAWAY

Understanding sentiment through groups of sparse features

Small groups reveal sentiment information.

Good prediction + Similar groups across masks

A separate BERT test finds a small effect from editing the features.

Next steps: separate number from formatting, separate syntax and role labels from templates, and improve arithmetic recovery.

github.com/misterpawan/manifold-sae-paper.git

aiwranglers.org optimalstar.org

CODE REPOSITORY

Currently private.

- Small feature groups help us read sentiment, and many selected features remain relevant across masking levels.
- A separate BERT experiment shows a small effect when we edit the selected group.
- The framework also guides tests of number, syntax, roles, and arithmetic. Thank you.