The Geometry of Metaphor: Probing and Steering Metaphor Representations in Large Language Models

Shihui Li*, **Mette Andersen***, **Martha Lewis**
Institute for Logic, Language, and Computation
University of Amsterdam
shihui.li@student.uva.nl, {m.f.andersen, m.a.f.lewis}@uva.nl

Abstract

Metaphor is a crucial aspect of human cognition. As such, we would like to understand how it is represented in large language models (LLMs) and how it might be controlled. We use mechanistic interpretability techniques to identify causal subspaces in a range of LLMs covering different architectures and sizes. We show that we are able to steer output on classification tasks across different parts of speech. We are also able to steer output on generation tasks, indicating that these techniques can be used to control the metaphoricity of LLM output.

1 Introduction

Metaphor is an essential part of everyday human language use. Sources such as Steen et al. (2010); Shutova (2010) show that metaphor pervades language. Furthermore, metaphor forms a basis for the fundamental creativity and concept formation that we are capable of (Lakoff and Johnson, 2008). Large language models such as Llama 3 (Grattafiori et al., 2024) have some capacity to understand and produce metaphorical language, although whether they are up to human performance is debated (Chakrabarty et al., 2021; Stowe et al., 2022; Ichien et al., 2024).

One way to understand the capacities of large language models is via *mechanistic interpretability*: by examining the internals of large language models via, for example, the Logit Lens[1], causal interventions (Meng et al., 2022), or circuit discovery (Conmy et al., 2023). Sharkey et al. (2025) provides a thorough review and discussion of open problems in the area. One recent paper, Marks and Tegmark (2024), argued that Llama models exhibited a 'truth direction' at particular layers, which was common to different datasets, and that the models' output could be steered by causal intervention on this layer.

Finding such a direction for metaphor would give insight into how models represent metaphor. This has been investigated to some extent (Boisson et al., 2024) but not yet at the mechanistic level. Finding this direction could also provide a useful 'dial' that could potentially be used to steer the metaphoricity or literalness of output text, using techniques from e.g. Li et al. (2023). In this paper, we use activation patching (Meng et al., 2022) and distributed alignment search (Geiger et al., 2024) to find a 'metaphoricity direction' or subspace within a given model. We show that these directions are causally effective: we are able to change the output of models in classification tasks for metaphor identification and interpretation; and in a generation task we are able to steer the model towards more metaphorical output.

Contributions

- We identify directions and subspaces for metaphoricity in a range of models, and show that these have some causal effect.
- We show that these directions are somewhat consistent across different metaphor datasets.
- We test seven models from four different model families (Gemma (Riviere et al., 2024), Qwen (Bai et al., 2023), Llama (Grattafiori et al., 2024), Mistral (Jiang et al., 2023)) across small (7-9B) and medium-large (27-70B) sizes.
- We show that steering models in a generation task can change model output from more literal to more metaphorical.

*Joint first authors

[1] https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens

Proceedings of the 9th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 965–985
October 29, 2026 ©2026 Association for Computational Linguistics

2 Background and Related work

Metaphor identification and interpretation In metaphor processing, research has focussed on two main tasks: *metaphor identification*, that is, identifying whether a word is metaphorical, and *metaphor interpretation*, that is, identifying the meaning of a metaphorical utterance. Work in using LLMs for metaphor identification include Fuoli et al. (2026); Reimann and Scheffler (2025). On the metaphor interpretation side, there has been a range of work showing that modern LLMs have relatively strong performance at understanding metaphor, especially when fine-tuned (Ichien et al., 2024; Liu et al., 2022b). Since metaphor is a key part of human language, we would like to understand how LLMs interpret and identify metaphor.

Activation Patching Mechanistic interpretability techniques attempt to understand and affect the output of large language models by inspecting their internal representations, for example by looking at the vectors of activations produced at each layer of the model (Meng et al., 2022; Marks and Tegmark, 2024). A common technique is to then *causally intervene* on these representations. This means that we examine how a model's output changes when we alter its internal representations, for example by *activation patching*, introduced in Meng et al. (2022); that is, we save the activations for one input, and patch these into the activations for another input, examining which layers and tokens affect the output.

In Marks and Tegmark (2024), the authors use activation patching to find a combination of layer and token input that represents a concept of truth in a language model.

Distributed Alignment Search (DAS) (Geiger et al., 2024) is a method for finding causal concepts in model activations, based on the assumption that full-vector activations may be polysemantic, i.e. may be activated for different concepts. The method has been used in existing benchmarks (Mueller et al., 2025) for causal variable localization, and compared to other methods (SAEs, PCA, full-vector), DAS has been shown to find concepts that achieve superior causal performance. In this paper, we use DAS to find a causal concept of metaphoricity in the model's residual stream, and compare the performance of steering using DAS with the existing mass-mean steering used by Marks and Tegmark (2024).

Controlling Generation Controlling the style of LLM generated text is also an active area of research. Techniques used include training a separate model (Liu et al., 2022a), or prompt-based style control such as in (Liu et al., 2024). Relatedly, Li et al. (2023) alter the truthfulness of generated text via inference-time interventions.

Summary Metaphor identification, interpretation, and generation are key aspects of language use. So far, in LLMs, these abilities have mostly been investigated by looking at behavioural outputs. In this paper, we use activation patching and distributed alignment search to investigate firstly whether metaphorical and literal sentences are linearly separable in activation space, secondly whether we can steer models towards literal or metaphoric output, and lastly whether the steering vectors are common to different datasets. We control the metaphoricity of generated output using the inference-time intervention technique introduced by Li et al. (2023). To our knowledge, this is the first attempt to control the metaphoricity of LLM output via intervention on the activations, rather than via prompting or building an external model.

3 Datasets and Tasks

3.1 Datasets

We use three datasets covering two different parts of speech (nouns and adjectives)[2]. Using mechanistic interpretability techniques requires that models are able to perform to a high standard on the task, and the structure of the dataset also needs to be fairly controlled. We therefore use relatively simply structured datasets. Following Marks and Tegmark (2024), we use datasets with the following properties: 1) each sentence contains exactly one metaphorical target word; 2) sentences are short and structurally simple; 3) within each dataset, sentences have the same syntactic structure.

We use the following three datasets. Examples of sentences with metaphorical nouns and adjectives are shown in Table 1.

Noun dataset (NOUN). We start from the small training set of Fig-QA (Liu et al., 2022b), which contains 200 human-written creative metaphors. Each example is paired with one correct rephrase and one incorrect rephrase in a Winograd-style

[2]We initially also used a verb dataset, however model performance on this dataset was not sufficiently high to probe with interpretability techniques.

Category	Example sentence
Noun	The book is a puzzle.
Adj	The welcome is warm.

Table 1: Example sentences with metaphorically used nouns and adjectives. In example sentences, metaphorical target words are coloured blue.

format. Every two metaphor examples form a pair representing the meaning of a particular rephrase.

We first remove idioms, multi-clause descriptions, and long narrative sentences (e.g., *"The painting has the creativity of Rembrandt and Picasso getting their paintings married"*). We then use ChatGPT (OpenAI, 2025) to rewrite the remaining examples into short, direct metaphors with the parallel structure *"The noun$_1$ is (a/an) noun$_2$"*. For example, the original sentence *"The bear was the size of a whale after eating a smaller whale"* is simplified to *"The bear is a whale"*. In each sentence, noun$_2$ metaphorically describes noun$_1$. We manually verify that each sentence is grammatically correct and preserves the original metaphorical mappings. The final dataset contains 130 sentence tuples. In the rest of the paper we refer to this dataset as **NOUN**.

Adjective dataset (ADJ). We use the ADJ+N metaphor/literal dataset proposed by Tsvetkov et al. (2014). The test set contains 100 adjective–noun pairs with both literal and metaphorical usages. We modify this dataset to have a sentence format using the template *"The N is ADJ."* In this dataset, the adjective carries the metaphorical meaning and modifies the noun. In the remainder of this paper, we refer to this dataset as **ADJ**.

Domain-specific adjective dataset (DOM). We also use the dataset by Gutiérrez et al. (2016) for DAS training, steering prediction and steering generation. The dataset consists of adjective-noun pairs across 8 different domains (clarity, depth, light, strength, substance, taste, temperature, texture). Each sample is labelled literal (e.g. "warm coffee") or metaphorical (e.g. "warm welcome"). We evaluate on n=100 per steering direction from the held-out set not used for training DAS. In the remainder of this paper we refer to this dataset as **DOM**.

3.2 Tasks

We carry out an **identification** task, an **interpretation** task, and a **generation** task.

Identification Models are prompted with a four-shot setup to label a sentence as either 'literal' or 'metaphoric'. The adjective datasets from Tsvetkov et al. (2014) and Gutiérrez et al. (2016) are in the form `adjective noun`. We rephrase these into sentences of the form 'The `noun` is `adjective`'. The full prompt for **ADJ** is then as follows. See Appendix A for the prompt for **NOUN**

The town is sleepy. This sentence is: metaphoric
The sand is white. This sentence is: literal
The moon is young. This sentence is: metaphoric
The banana is ripe. This sentence is: literal
`sentence`. This sentence is:

Interpretation Models are prompted with a four-shot setup. In each shot, a metaphorical sentence and a candidate paraphrase are given, and the paraphrase is labelled as 'apt' or 'inapt'. This task uses **NOUN**, which is based on FigQA (Liu et al., 2022b). Apt/inapt pairs are taken from their setup. The full prompt is given below:

Determine whether the paraphrase is apt or inapt.
The politician is glass: The politician is transparent. This paraphrase is: apt.
The politician is a wall: The politician is transparent. This paraphrase is: inapt.
The cake is heaven: The cake is not delicious. This paraphrase is: inapt.
The cake is heaven: The cake is delicious. This paraphrase is: apt.
`sentence`: `rephrase`. This paraphrase is:

Generation Models are prompted with a five-shot setup to output an interpretation of a sentence that could be either literal or metaphorical. The dataset used is **DOM**, and we reword adjective-noun pairs into sentences of the form 'The `noun` was `adjective`'. The full prompt is given below:

The welcome was warm. The meaning of 'warm' is: affectionate and emotionally inviting.
The lamp was bright. The meaning of 'bright' is: emitting a strong, clear light.

The novel was deep. The meaning of 'deep' is: intellectually complex and rich in meaning.
The stone was smooth. The meaning of 'smooth' is: having an even surface with no roughness.
The defeat was bitter. The meaning of 'bitter' is: painful, hard to accept, and leaving resentment.
The noun was adj. The meaning of 'adj' is:

4 Methods

4.1 Identifying areas encoding metaphoricity

For each of the datasets and tasks described above, we use one of two techniques designed to isolate the part of the model used to encode metaphoricity. For the identification and interpretation tasks across **NOUN**, **ADJ**, and **DOM**, we use **activation patching**. For **DOM** and for the generation task, we also use **distributed alignment search (DAS)** to identify a linear subspace encoding metaphoricity. We use the **mass-mean probe** from Marks and Tegmark (2024) and the learned DAS subspaces to steer model output in the identification and interpretation tasks, requiring single-token output. Finally, we use the mass-mean probe and DAS subspaces to steer generation over multiple tokens, finding that it is possible to steer models towards more metaphorical or more literal output.

4.1.1 Activation Patching

In the identification task, we use the following procedure. On **ADJ**, we insert the following sentence into the prompt template in Section 3.2 to form a prompt p_L= *The skin is dry.* For prompt p_L we expect the output 'literal'. We then form a paired prompt p_M replacing *skin* with *wit*. For this prompt we expect the output 'metaphoric'. To find out which hidden states contain information about the metaphoricity of the sentence, we run the model on the input p_M and cache the activations $h_{i,\ell}(p_M)$ for each token position i and layer ℓ. Then, for each i and ℓ, we run the model on P_L and modify the model's forward pass by swapping $h_{i,\ell}(P_L)$ for $h_{i,\ell}(P_M)$. Finally, we record the difference in log probability between the tokens "metaph"[3] and "literal". The larger the difference, the more influential the hidden state in position i and layer ℓ on the model's prediction. For the dataset **NOUN**, we use the prompt pair p_L=*The breeze is cold* and p_M=*The breeze is ice*. For the interpretation task, we use the prompt pair p_{inapt} = *The puzzle is rocket science: The puzzle is easy* and p_{apt} = *The puzzle is rocket science: The puzzle is difficult*.

[3]LlaMA-3.1-8B tokenizes "metaphor" into subwords "metaph" and "or". In our study, we examine the model's first output token "metaph".

4.1.2 Distributed Alignment Search

We use Distributed Alignment Search (DAS) (Geiger et al., 2024) to localize, if it exists, a low-dimensional, linear subspace encoding metaphoricity. After identifying the most causally implicated token-position in the residual stream using full-vector activation patching, DAS trains a rotation matrix on top of these activations. The rotation matrix is trained to maximize Interchange Intervention Accuracy (IIA), which is the fraction of samples where patching flips model prediction in the expected direction.

Concretely, let $\mathbf{h} \in \mathbb{R}^d$ denote the residual stream at a fixed layer ℓ and token position p. As will be seen in Figure 3, the full stop token at the end of the target sentence in each prompt is highly causally implicated. We perform DAS at this token at the layer that optimises the IIA score.

DAS learns a matrix $\mathbf{R} \in \mathbb{R}^{d\times k}$ with orthonormal columns ($\mathbf{R}^\top\mathbf{R} = \mathbb{I}_k$), defining a k-dimensional subspace of the representation space. Given a source hidden state $\mathbf{h}^{\text{source}}$ and a base hidden state $\mathbf{h}^{\text{base}}$, the intervention replaces the subspace component of $\mathbf{h}^{\text{base}}$ with that of $\mathbf{h}^{\text{source}}$:

$$\mathbf{h}_{\text{patched}} = \underbrace{\mathbf{h}}_{\substack{\text{original}\\\text{activation}}} - \underbrace{\mathbf{R}\mathbf{R}^\top\mathbf{h}}_{\substack{\text{remove original}\\\text{subsp. coords}}} + \underbrace{\mathbf{R}\mathbf{R}^\top\boldsymbol{\mu}_{\text{target}}}_{\substack{\text{inject target}\\\text{subsp. coords}}}$$

During training, we alternate intervention directions: on even steps we patch literal representations into metaphor prompts (aiming to flip the output to literal), and on odd steps the reverse. The loss for a single example is $\mathcal{L} = \text{logit(other)} - \text{logit(target)}$ where target is the correct post-intervention label and other is the original label. $\mathbf{R}$ is updated via Adam for 1,000 steps while the model weights are frozen.

IIA is the fraction of samples that are correctly flipped. In our case, if $\mathbf{h}^{\text{base}}$ is predicted as "metaphor" for the model, it should flip to "literal" post-patching to contribute positively to IIA. Hence, IIA = 1.0 means we have found the variable capable of flipping *all* held-out examples.

4.2 Model steering

We evaluate whether we can steer incorrect predictions towards correct ones using two methods: full-vector mass-mean steering, as used by Marks and Tegmark (2024), and steering using the mean activations for metaphor-sample and literal-sample activations in the DAS subspace. The means are computed from correctly predicted samples, and we steer incorrectly predicted samples in both directions.

Mass-mean steering is implemented as:

$$\mathbf{h}_{\text{steered}} \leftarrow \mathbf{h} + \alpha(\boldsymbol{\mu}_{\text{met}} - \boldsymbol{\mu}_{\text{lit}}) \quad (1)$$

DAS steering is implemented as:.

$$\mathbf{h}_{\text{steered}} \leftarrow \mathbf{h} - \alpha\,\mathbf{R}\mathbf{R}^{\top}\mathbf{h} + \alpha\,\mathbf{R}\mathbf{R}^{\top}\boldsymbol{\mu}_{\text{target}} \quad (2)$$

We set $\alpha = 1$ for the identification and interpretation tasks on **NOUN**, **ADJ**, and **DOM**. On **DOM** we train DAS matrices firstly across the whole dataset, and subsequently for each individual domain.

4.2.1 Generation of multiple tokens

To test whether the DAS subspace encodes a general-purpose concept of metaphor, which extends beyond the prediction task it was trained on, we steer free-form generation. We prompt the model as described in section 3.2. During decoding, we steer every generated token as in Li et al. (2023): we use a hook function to save the activation we use for patching, and inject it at every generated token.

We compare baseline generation, mass mean, and DAS outputs across $\alpha \in \{1, 2, 5, 10\}$ and both steering directions (lit $\rightarrow$ met and met $\rightarrow$ lit).

Evaluation using LLM-as-judge To quantify the degree to which generated output changed from explaining a literal use of an adjective to a metaphorical use and vice versa, we use Qwen 2.5 72B Instruct as a judge. Only the first sentence of each completion is judged, and a degeneracy filter applied before judging: a completion is marked degenerate if any single word repeats ≥ 3 times. The judge is prompted using the prompt shown in Appendix K.

Metrics We report the following metrics: **Flip rate (MM / DAS)**: fraction of of completions where judge label matches the target direction. Degenerate responses are counted as 0, and we only consider the samples where the judge was able to judge correctly on the baseline generation pre-steering.

4.3 Models Evaluated

We evaluate seven instruction-tuned models across four model families[4] on metaphor identification and interpretation tasks. These are: Gemma 2 9B IT; LLaMA 3.1 8B IT; LLaMA 3.1 70B IT; Mistral 7B IT; Mixtral 8×7B IT; Qwen 2.5 7B IT; Qwen 2.5 72B IT, where IT stands for intstruction tuned.

5 Results

5.1 Baseline Performance

We evaluate all models on the identification and interpretation tasks across all datasets (see Table 2). We can see that all models are able to perform these tasks to a reasonable standard with the exception of Gemma 27B on **DOM**. This modes is excluded from further analysis on this dataset.

5.2 Identifying Metaphoricity

5.2.1 Activation patching

Examples of patching from the interpretation and identification tasks are presented in Figure 1. Additional patching results are shown in Figures 3 and 4 in the Appendix.

Across both parts of speech and both tasks, we observe three groups of activations that most strongly influence the output (Figure 1). Two of these groups are expected. Group (a) corresponds to tokens that differ between the patching source and base prompts (e.g., “skin” and “wit”). Group (c) corresponds to the final token “:”, which is natural because the hidden states of the final token directly determine the next-token prediction.

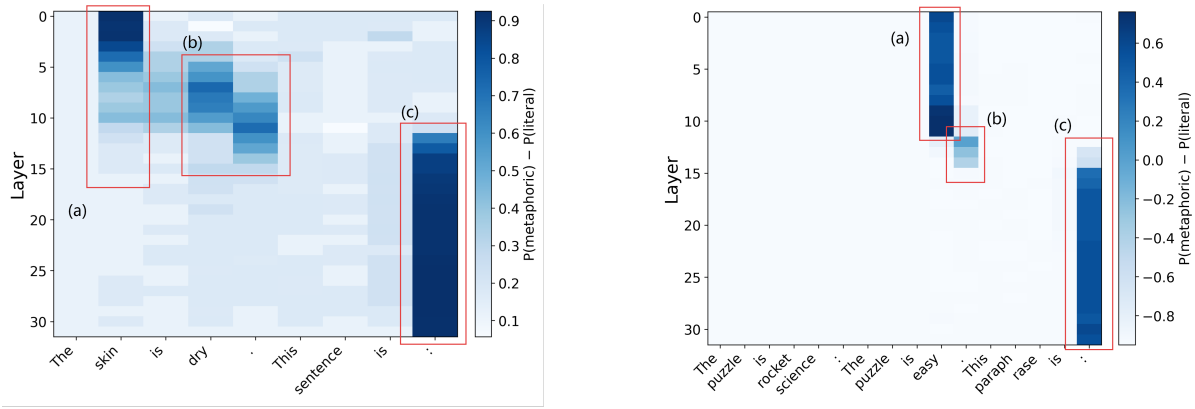
Group (b) contains the final token “.” in the sentence or paraphrase, as well as the token immediately preceding it. We hypothesize that Group (b) activations encode information about the metaphoricity of the sentence or paraphrase.

Based on the patching results, we extract the hidden states of the selected period token on the layer which has the most influence on the model’s output (e.g. 11th layer for Llama-3.1-8B, 29th layer for Llama-3.1-70B). We visualize LLM representations of our datasets using principal component

[4] Where available, we also evaluated non-instruction tuned models, but found that their performance on the tasks was not strong enough for the application of interpretability techniques.

Task	Dataset	LLaMA-3.1-8B	LLaMA-3.1-70B	Qwen-2.5-7B	Qwen-2.5-72B	Mistral-7B	Mixtral-8x7B	Gemma-2-9B	Gemma-2-27B
Identification	**ADJ**	0.755	0.915	0.860	0.930	0.760	0.870	0.930	0.915
	NOUN	0.840	0.905	0.815	0.888	0.645	0.881	0.873	0.900
	DOM	0.841	0.907	0.805	0.911	0.814	0.825	0.881	0.464
Interpretation	**NOUN**	0.727	0.862	0.696	0.815	0.765	0.796	0.845	0.858

Table 2: Baseline accuracy on the 4-shot evaluation on identification and interpretation tasks across all models and datasets. DOM results use the M&T 5-shot prompt. Gemma-2-27B excluded from **DOM** analysis.



(a) Llama 8b *identification* task

(b) Llama 8b *interpretation* task

Figure 1: **Activation patching results for Llama 3.1 8B. (a) Identification task.** The prompts use the literal sentence *"The skin is dry."* and the metaphorical sentence *"The wit is dry."* from **ADJ** We patch activations from the literal prompt into the metaphorical prompt. **(b) Interpretation task.** The prompts use the metaphorical sentence *"The puzzle is rocket science."*, the correct paraphrase *"The puzzle is easy."*, and the incorrect paraphrase *"The puzzle is difficult."* from **NOUN**. We patch activations from the apt paraphrase into the inapt paraphrase.

analysis (PCA). As shown in Figure 2, the first two principal components (PC1 and PC2) linearly separate metaphorical from literal and apt from inapt across **NOUN** and **ADJ** and both tasks. More results are available in Figures 5,6 and 7 in Appendix C.

5.2.2 DAS training

Table 3: Global (all domains pooled) DAS IIA across models. IIA reported as mean ± std with 95% CI.

Model	ℓ	k	IIA ± std	95% CI
Gemma-9B	19	16	0.807 ± 0.007	$[0.799,\ 0.815]$
LLaMA-8B	11	16	0.793 ± 0.006	$[0.783,\ 0.802]$
Mistral-7B	10	32	0.711 ± 0.019	$[0.698,\ 0.722]$
Mixtral-8x7B	11	16	0.691 ± 0.026	$[0.680,\ 0.701]$
Qwen-7B	14	16	0.830 ± 0.000	$[0.821,\ 0.839]$
LLaMA-70B	28	32	0.640 ± 0.025	$[0.630,\ 0.650]$
Qwen-72B	52	32	0.843 ± 0.024	$[0.836,\ 0.850]$

We select the optimal layer for training DAS using plain activation patching at the full stop token across layers. We extract the full hidden vector at this token at the optimal layer, and train two DAS matrices (k=16 and k=32). In Table 3 we report best layer ℓ, subspace dimension k and IIA score.

We are interested in whether the model has a domain-general causal encoding of metaphor, or multiple domain-specific ones. We therefore compare the IIA for a global DAS matrix and multiple domain-specific DAS matrices on a held out evaluation set for each domain (not used for training global DAS, nor the domain-specific variant). Results for Llama 3.1 8B are reported in Table 4. We see that the global DAS matrices have higher IIA than domain-specific matrices. For all results, see Table 9 and Table 10.

Table 4: LLaMA-8B: Domain-specific DAS IIA vs. global DAS transfer on held-out evaluation set

Concept	Domain-specific DAS	Global DAS (transfer)
Clarity	0.316 ± 0.070 $[0.262, 0.369]$	0.500 $[0.404, 0.596]$
Depth	0.343 ± 0.010 $[0.299, 0.386]$	0.910 $[0.868, 0.952]$
Light	0.322 ± 0.122 $[0.294, 0.352]$	0.727 $[0.674, 0.776]$
Streng.	0.606 ± 0.061 $[0.582, 0.628]$	0.967 $[0.953, 0.982]$
Sub.	0.309 ± 0.059 $[0.248, 0.371]$	0.814 $[0.714, 0.900]$
Taste	0.645 ± 0.054 $[0.588, 0.706]$	0.842 $[0.763, 0.921]$
Temp.	0.470 ± 0.051 $[0.435, 0.506]$	0.568 $[0.500, 0.627]$
Tex.	0.343 ± 0.022 $[0.311, 0.376]$	0.713 $[0.657, 0.765]$

5.3 Model steering

5.3.1 Identification and Interpretation

On the identification and interpretation tasks on **NOUN** and **ADJ**, we intervene on model outputs

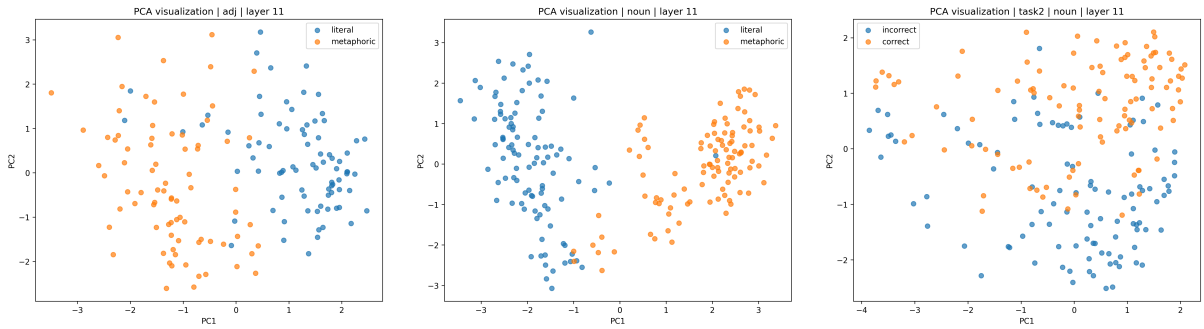


Figure 2: PCA visualization of activations at layer 11 for Llama 3.1 8B, for the identification task on the adjective (left) and noun(middle) datasets, and for the interpretation task on the adjective dataset (right). Notice that literal and metaphoric sentences are more or less linearly separable.

Task	train/test	probe	met→lit /apt→inapt	lit→met /inapt→apt
Task 1	Adj/Adj	LogReg	.038	.051
		MM-diff	**.227**	**.471**
	Noun/Noun	LogReg	.025	.041
		MM-diff	**.195**	**.449**
	Adj/Noun	LogReg	.026	.032
		MM-diff	**.138**	**.254**
	Noun/Adj	LogReg	.028	.040
		MM-diff	**.230**	**.666**
Task 2	Noun/Noun	LogReg	.011	.014
		MM-diff	**.355**	**.564**

Table 5: Normalized indirect effects (NIEs) of mass-mean difference (MM-diff) and logistic regression (LogReg) probes on the *identification* and *interpretation* tasks for LLaMA-3.1-8B. Higher NIE values indicate stronger causal steering effects.

as in Eq. (1). Then, we construct a linear probe using the direction $\boldsymbol{\theta} = \boldsymbol{\mu}_{\text{met}} - \boldsymbol{\mu}_{\text{lit}}$:

$$p(\text{metaphor} \mid s) = \sigma\big(\boldsymbol{\theta}^\top \mathbf{h}(s)\big),$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$. For the *interpretation* task, we define directions for apt/inapt pairs using the corresponding means.

First, we use the mass-mean probe to flip the true positive outputs in test set into the opposite label. We then compare MM probes with standard logistic regression (LR) probes trained on the same representations. For both probes, we calculate the normalized indirect effect (NIE) of the intervention. For details of the NIE calculation, see Appendix J. Results for Llama-3.1-8B are reported in Table 5. Results across all models are in Appendix J.

5.3.2 Mass-mean and DAS steering

We compare the performance of mass-mean steering and subspace steering using the global DAS matrices trained in Section 5.2.2. We choose initially incorrect predictions on the identification task on the **DOM** dataset. We use the two steering methods to steer output to the correct prediction. Results are reported in Table 6. We see that we are able to achieve fairly high performance across models. For a full overview, see Table 11 in Appendix F.

Model	MM flip rate	DAS flip rate
Gemma 2 9B	75.0%	**86.0%**
LLaMA 3.1 8B	54.0%	**71.0%**
LLaMA 3.1 70B	57.0%	**65.0%**
Mistral 7B	**64.0%**	45.0%
Mixtral 8×7B	**72.0%**	61.0%
Qwen 2.5 7B	**62.0%**	61.0%
Qwen 2.5 72B	88.0%	**91.0%**

Table 6: Repair rates on initially-incorrect items. **MM flip rate**: proportion of incorrect items flipped to the correct answer by full-residual-stream patching. **DAS flip rate**: proportion flipped by intervening only on the learned subspace. Bold marks the higher value per row.

5.3.3 Generation

Whilst there are differences between models, we find that $\alpha = 5$ enables the most effective steering on average. Results for Llama-3.1-8B are presented in Table 7. For a full overview, we refer to Tables 12 and 13 in Appendix G.

To validate the results of the LLM-as-judge, we qualitatively inspect the generated outputs. We display some generations in Table 8. More examples, including successful and unsuccessful generations, are included in Table 14 in the Appendix.

6 Discussion and Conclusions

Our results show that there are clear directions and subspaces associated with metaphoricity across a

	Lit → met		Met → lit	
Model	**MM**	**DAS**	**MM**	**DAS**
Gemma 2 9B	0.33	0.12	0.25	0.12
LLaMA 3.1 8B	0.44	0.34	0.03	0.09
LLaMA 3.1 70B	0.34	0.28	0.15	0.16
Mistral 7B	0.52	0.19	0.28	0.10
Mixtral 8×7B	0.38	0.16	0.24	0.11
Qwen 2.5 7B	0.26	0.47	0.21	0.12
Qwen 2.5 72B	0.29	0.15	0.08	0.19

Table 7: Flip rates at $\alpha = 5$ for both intervention methods and both directions.

range of models and datasets (Figure 2 and Figures 5, 6 in the Appendix). We are able to train DAS matrices which effectively alter output on a classification task (Table 3), and we see that training a global metaphor subspace is more effective than training domain-specific subspaces (Table 4), indicating that we have learnt a general concept of metaphor. These directions are moreover causally effective: using the mass-mean probe we are able to steer output in classification tasks, and this steering is more effective than steering using the direction identified by logistic regression (Table 5). We do note that the NIE values are somewhat lower than in (Marks and Tegmark, 2024), however, it is also the case that metaphor identification and interpretation is a more difficult task than judging truth and falsity. We also see that there are higher NIEs steering from literal to metaphor than vice-versa. Creating a literal meaning from a metaphorical meaning is perhaps conceptually more difficult: *warm coffee* can readily be interpreted metaphorically via personification of the coffee, whereas *warm welcome* is more difficult to interpret literally.

Table 8: Example generations for LLaMA-8B under steering at $\alpha = 5$. Baseline = unintervened model, MM = mass-mean steering, DAS = DAS intervention. Both examples: true label *literal*, target *metaphorical*.

Method	Judge prediction	Generation
adj-noun pair: `icy pavement`		
Baseline	literal	extremely cold and slippery.
MM	metaphorical	extremely cold and unfriendly.
DAS	metaphorical	extremely cold and unfriendly.
adj-noun-pair: `heated washer`		
Baseline	literal	having a high temperature.
MM	metaphorical	having a high temperature, often due to strong emotions.
DAS	metaphorical	having a strong, intense, or passionate quality.

We further steer generation to more metaphorical output. In Table 7, we see that the mass-mean probes are generally more effective than DAS steering, although this differs across model. Example generations are given in Table 8. In future work, this technique could be developed to give a way of controlling the metaphoricity of generated output in a more direct way than prompting.

We acknowledge that IIA scores are moderate and variable, and indeed this does imply that a proportion of the metaphoricity information is captured elsewhere. We note however, that previous work which has used DAS for locating concepts in model activation achieve similar scores (see e.g. (Mueller et al., 2025)). Achieving a perfect IIA score is not necessarily achievable in practice (Geiger et al., 2024), and this is a limitation of using DAS for concept-finding.

On the basis of Table 4, we concluded that the DAS matrix trained on all domains captures a general concept of metaphoricity, as evidenced by its stronger transfer IIA results compared to the results achieved with its domain-specific counterpart. One confound could be the training data size: Compared to the domain-specific matrices trained on n samples, the global DAS matrix is trained on those *and all the training samples from all other domains*. As a control, we train a matched-n global DAS matrix. This matrix trained on the number of samples in each domain corresponding to the number of samples from the smallest training domain ($n = 255$). We see in Appendix Table 15 that for Gemma 2 9B, training a matched-n global DAS matrix still achieves a better or equal transfer IIA score in 6/8 transfer domains.

For our generation experiment, we carried out a human validation of the LLM-as-judge. For the Gemma 9B model outputs, one of the authors of this paper has annotated 15% (100 items) as either metaphor, literal, or incorrect (nonsensical or incorrect output). Raw agreement is 0.79, Cohen's kappa is 0.60, indicating moderate to substantial agreement. This indicates that the judge annotations are mostly reliable. This suggests that the concept we have isolated in the classification task is at least partially sufficient to flip free-form generation. However, generally, we do achieve substantially lower IIA scores for the generation task compared to the classification task. Our DAS concept allows us to steer model classification to a moderately high degree, but this effect is less amplified

for generation, and requires that we scale this signal by $\alpha = 5$. There can be different reasons for this. It is possible that the metaphoricity concept we isolate in the classification task is functionally adapted to this task. The model *could* have another metaphoricity concept located in a different position adapted for generation. On the other hand, the fact that the concept has some meaning-steering effect suggests that we have captured a common concept of metaphoricity, at least partially.

Further, we employed a controlled, templated setup for measuring metaphoricity in LLMs. We made this choice to effectively isolate the parts of the prompt that have a causal effect on the output. The structured sentence patterns allowed us to localize causal effects to specific token positions. Therefore, we deliberately did not test representations of metaphoricity in a more naturalistic context (e.g. poems), limiting the external validity of our findings.

6.1 Conclusions

We have given evidence that metaphoricity is encoded as an interpretable direction in a variety of LLMs, illustrated as linear separability in PCA plots. We intervene on model activations and show that these directions are causally effective in changing model outputs in both classification and generation tasks. Future work in this area could extend to other aspects of writing style: at present, much work is based in prompting, and these techniques may give more control over style modifications.

7 Limitations and Risks

- This work is carried out only on English metaphor. Linear separability and interpretable directions could be very different in other languages.
- Mechanistic interpretability work requires very controlled settings and datasets, whereas metaphor can extend across multiple sentences, paragraphs, and discourse. The datasets we use are therefore only very simplistic uses of metaphor. More complex metaphor use is an area for future research.
- We cannot carry out our experiments on the most highly performing closed models, since they do not have open weights, and computational resources would preclude this.
- **Risks** In its present form, risks resulting from this work are fairly low. However, model steering techniques can potentially be risky. It is possible to steer towards undesirable as well as desirable behaviour.

Acknowledgements

This work was carried out using the National Supercomputer Snellius made available via SURF (www.surf.nl) and the University of Amsterdam. ML gratefully acknowledges financial support from the Sectorplan science fund of the Dutch government.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Joanne Boisson, Asahi Ushio, Hsuvas Borkakoty, Kiamehr Rezaee, Dimosthenis Antypas, Zara Siddique, Nina White, and Jose Camacho-Collados. 2024. How are metaphors processed by language models? the case of analogies. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 365–387.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Matteo Fuoli, Weihang Huang, Jeannette Littlemore, Sarah Turner, and Ellen Wilding. 2026. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning. *Applied Corpus Linguistics*, page 100204.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten,

Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

E. Dario Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics.

Nicholas Ichien, Dušan Stamenković, and Keith J Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 39(4):296–309.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*.

Ao Liu, An Wang, and Naoaki Okazaki. 2022a. Semi-supervised formality style transfer with consistency training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4689–4701, Dublin, Ireland. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022b. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-step: Controlling arbitrary style in text with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15285–15295.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:cs.AI*. Submitted on 10 Oct 2023 (v1), last revised 19 Aug 2024 (v3).

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, and 4 others. 2025. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

OpenAI. 2025. Chatgpt. `https://chat.openai.com`. Accessed: 2026-05-26.

Sebastian Reimann and Tatjana Scheffler. 2025. Using large language models to perform MIPVU-inspired automatic metaphor detection. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 10–21, Vienna, Austria. Association for Computational Linguistics.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, and 1 others. 2025. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037, Los Angeles, California. Association for Computational Linguistics.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. Metaphor in usage.

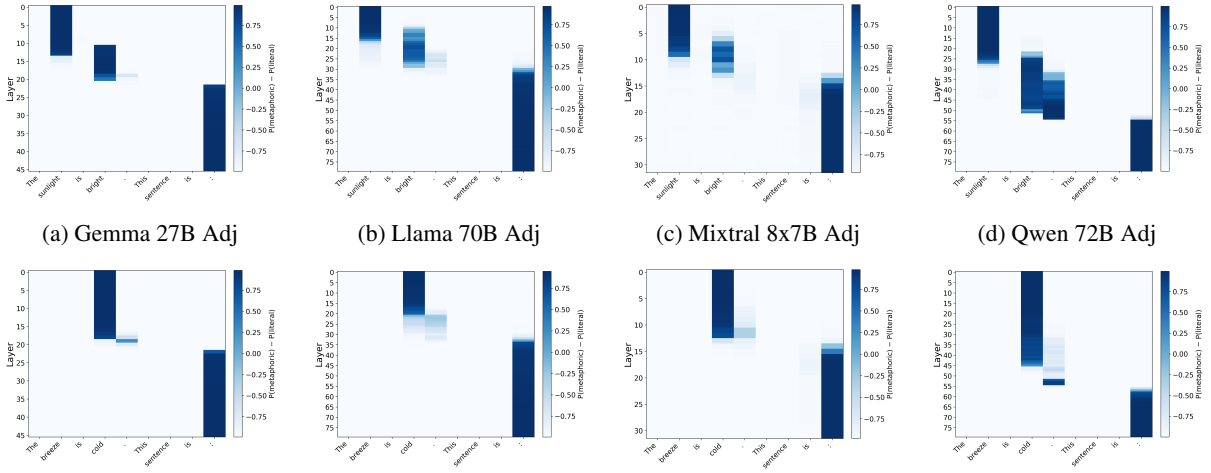
Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258.

A NOUN identification prompt

The politician is glass. This sentence is: metaphoric
The politician is transparent. This sentence is: literal
The cake is heaven. This sentence is: metaphoric
The cake is delicious. This sentence is: literal
`sentence`. This sentence is:

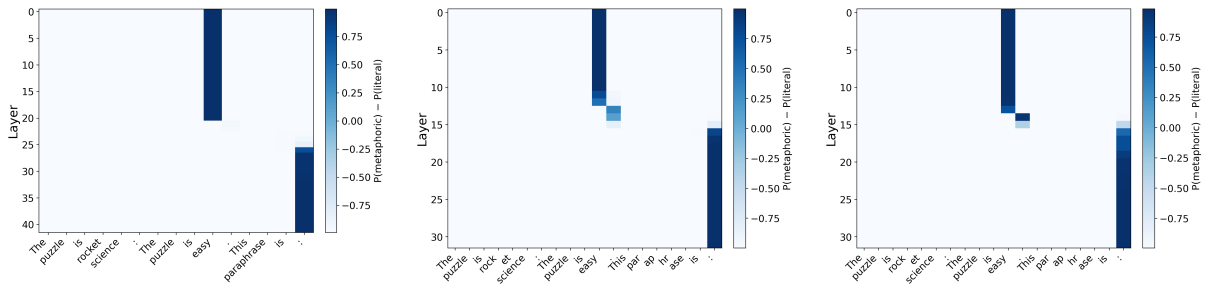
B Activation patching plots



(a) Gemma 27B Adj (b) Llama 70B Adj (c) Mixtral 8x7B Adj (d) Qwen 72B Adj

(e) Gemma 27B Noun (f) Llama 70B Noun (g) Mixtral 8x7B Noun (h) Qwen 72B Noun

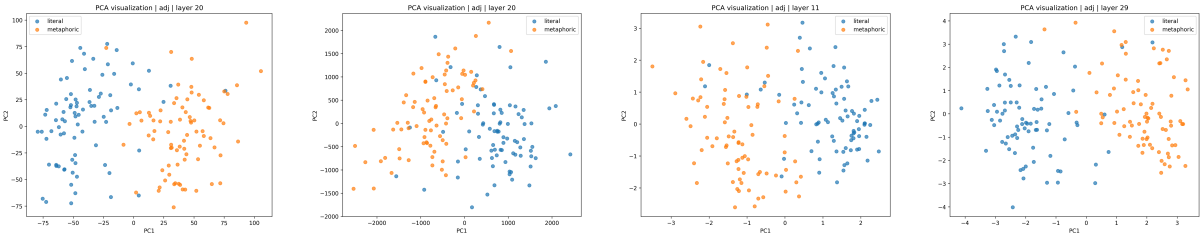
Figure 3: **Activation patching results for adjective and noun representations in the *identification* task.** Activations from the metaphorical prompt are patched into the literal prompt. **(a)–(d) Adjective dataset.** The prompts use the literal sentence *"The sunlight is bright."* and the metaphorical sentence *"The idea is bright."*. **(e)–(h) Noun dataset.** The prompts use the literal sentence *"The breeze is cold."* and the metaphorical sentence *"The breeze is ice."*.



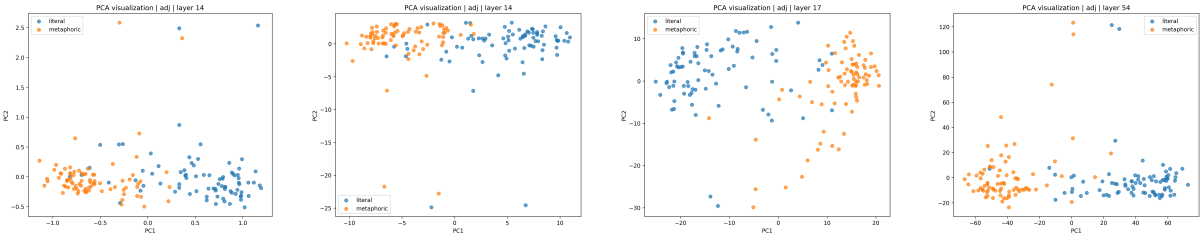
(a) Gemma 9B Noun (b) Mistral 7B Noun (c) Mixtral 8x7B Noun

Figure 4: **Activation patching results for noun representations in the *interpretation* task.** Activations from the inapt paraphrasing prompt are patched into the apt paraphrasing prompt. The prompts use the metaphorical sentence *"The puzzle is rocket science."*, the correct paraphrase *"The puzzle is difficult."*, and the incorrect paraphrase *"The puzzle is easy."*.

C PCA plots

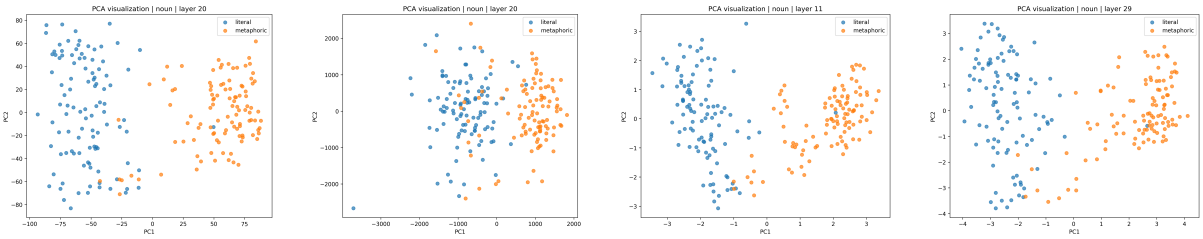


(a) Gemma-2 9B (b) Gemma-2 27B (c) Llama-3.1 8B (d) Llama-3.1 70B

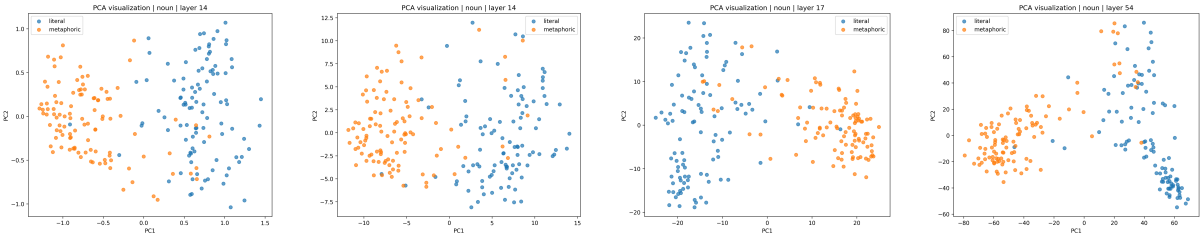


(e) Mistral 7B (f) Mixtral 8x7B (g) Qwen2.5 7B (h) Qwen2.5 72B

Figure 5: Identification task: PCA visualizations for adjective representations.

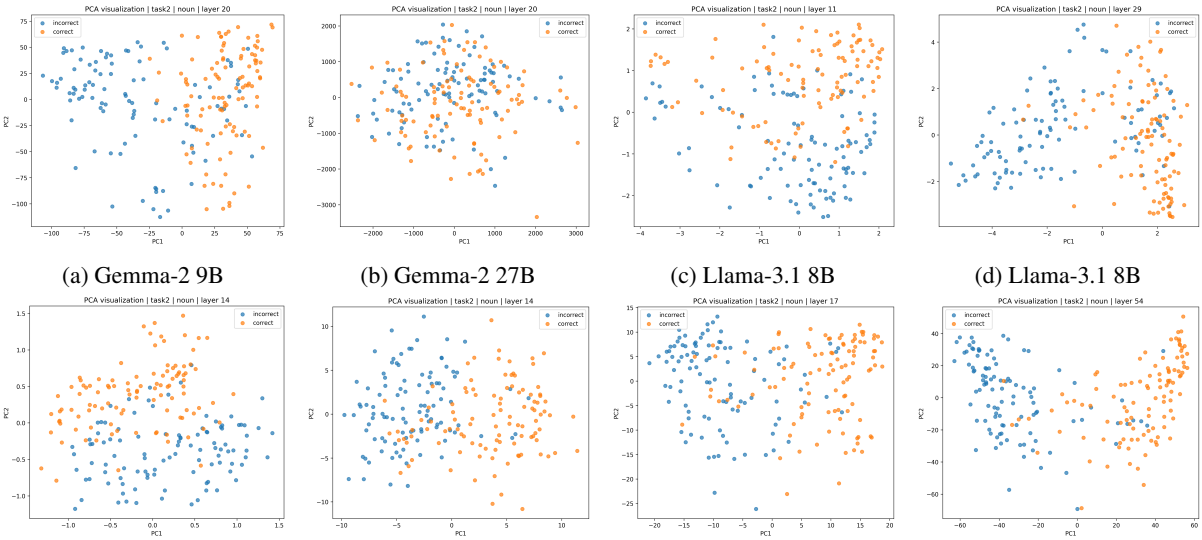


(a) Gemma-2 9B (b) Gemma-2 27B (c) Llama-3.1 8B (d) Llama-3.1 70B



(e) Mistral 7B (f) Mixtral 8x7B (g) Qwen2.5 7B (h) Qwen2.5 72B

Figure 6: Identification task: PCA visualizations for noun representations.



(a) Gemma-2 9B (b) Gemma-2 27B (c) Llama-3.1 8B (d) Llama-3.1 8B

(e) Mistral 7B (f) Mixtral 8x7B (g) Qwen2.5 7B (h) Qwen2.5 72B

Figure 7: Interpretation task: PCA visualizations for noun representations.

D Domain-specific DAS training at full-stop token position

Table 9: Domain-specific DAS IIA (mean $\pm$ std) across models. Best per domain in bold.

Domain	LLaMA-8B	LLaMA-70B	Gemma-9B	Mistral-7B	Mixtral-8x7B	Qwen-7B	Qwen-72B
Clarity	0.316 ± 0.070	0.351 ± 0.022	0.441 ± 0.010	0.245 ± 0.013	0.325 ± 0.026	$\mathbf{0.521 \pm 0.006}$	0.313 ± 0.040
Depth	0.343 ± 0.010	0.212 ± 0.095	0.227 ± 0.019	$\mathbf{0.447 \pm 0.008}$	0.406 ± 0.005	0.287 ± 0.061	0.079 ± 0.016
Light	0.322 ± 0.122	0.059 ± 0.006	0.524 ± 0.015	0.089 ± 0.008	0.158 ± 0.067	0.461 ± 0.017	$\mathbf{0.608 \pm 0.021}$
Strength	0.606 ± 0.061	0.508 ± 0.043	0.776 ± 0.011	0.417 ± 0.013	0.467 ± 0.033	$\mathbf{0.854 \pm 0.009}$	0.594 ± 0.087
Substance	0.309 ± 0.059	0.130 ± 0.037	0.500 ± 0.019	0.424 ± 0.013	0.319 ± 0.036	$\mathbf{0.514 \pm 0.000}$	0.062 ± 0.007
Taste	0.645 ± 0.054	0.248 ± 0.066	0.587 ± 0.043	0.662 ± 0.037	0.494 ± 0.008	0.796 ± 0.053	$\mathbf{0.826 \pm 0.009}$
Temperature	0.470 ± 0.051	0.077 ± 0.027	0.583 ± 0.021	0.542 ± 0.017	0.405 ± 0.014	0.672 ± 0.019	$\mathbf{0.838 \pm 0.054}$
Texture	0.343 ± 0.022	0.219 ± 0.084	0.506 ± 0.050	$\mathbf{0.810 \pm 0.035}$	0.613 ± 0.062	0.698 ± 0.014	0.533 ± 0.032

E Transfer: validation IIA using global DAS instead of domain-specific DAS

Table 10: Transfer IIA: global DAS evaluated on held-out domain (mean [95% CI]). Best per domain in bold.

Domain	LLaMA-8B	LLaMA-70B	Gemma-9B	Mistral-7B	Mixtral-8x7B	Qwen-7B	Qwen-72B
Clarity	$0.500\ [0.404, 0.596]$	$0.500\ [0.357, 0.625]$	$0.581\ [0.484, 0.661]$	$0.409\ [0.352, 0.466]$	$\mathbf{0.738}\ [\mathbf{0.680}, \mathbf{0.801}]$	$0.618\ [0.542, 0.701]$	$0.739\ [0.681, 0.792]$
Depth	$\mathbf{0.910}\ [\mathbf{0.868}, \mathbf{0.952}]$	$0.727\ [0.688, 0.767]$	$0.924\ [0.896, 0.952]$	$0.849\ [0.808, 0.887]$	$0.826\ [0.781, 0.868]$	$0.852\ [0.812, 0.893]$	$0.721\ [0.677, 0.767]$
Light	$0.727\ [0.674, 0.776]$	$0.240\ [0.192, 0.286]$	$0.580\ [0.523, 0.639]$	$0.455\ [0.400, 0.510]$	$\mathbf{0.821}\ [\mathbf{0.776}, \mathbf{0.864}]$	$0.627\ [0.571, 0.679]$	$0.763\ [0.717, 0.809]$
Strength	$\mathbf{0.967}\ [\mathbf{0.953}, \mathbf{0.982}]$	$0.956\ [0.938, 0.972]$	$0.919\ [0.895, 0.942]$	$0.595\ [0.545, 0.642]$	$0.777\ [0.734, 0.818]$	$0.933\ [0.908, 0.958]$	$0.959\ [0.940, 0.975]$
Substance	$\mathbf{0.814}\ [\mathbf{0.714}, \mathbf{0.900}]$	$0.469\ [0.344, 0.594]$	$0.643\ [0.500, 0.786]$	$0.600\ [0.486, 0.700]$	$0.586\ [0.471, 0.700]$	$0.757\ [0.657, 0.857]$	$0.771\ [0.671, 0.857]$
Taste	$0.842\ [0.763, 0.921]$	$0.656\ [0.544, 0.744]$	$0.830\ [0.750, 0.909]$	$0.714\ [0.600, 0.815]$	$0.517\ [0.400, 0.633]$	$0.871\ [0.774, 0.952]$	$\mathbf{0.977}\ [\mathbf{0.942}, \mathbf{1.000}]$
Temperature	$0.568\ [0.500, 0.627]$	$0.235\ [0.183, 0.287]$	$0.650\ [0.585, 0.705]$	$0.673\ [0.601, 0.744]$	$0.601\ [0.538, 0.664]$	$0.735\ [0.676, 0.790]$	$\mathbf{0.900}\ [\mathbf{0.863}, \mathbf{0.933}]$
Texture	$0.713\ [0.657, 0.765]$	$0.488\ [0.417, 0.554]$	$0.767\ [0.714, 0.813]$	$0.907\ [0.872, 0.943]$	$0.809\ [0.756, 0.858]$	$0.746\ [0.690, 0.798]$	$\mathbf{0.924}\ [\mathbf{0.890}, \mathbf{0.955}]$

F Steering incorrect predictions

Table 11: Flip rates on initially-wrong items: mass-mean steering vs. DAS intervention (mean [95% CI]). Best per row in bold.

Model	Wrong/Total	Mass-Mean flip	DAS flip
Gemma-9B	1140/8591	0.750 [0.660, 0.830]	**0.860 [0.790, 0.920]**
LLaMA-8B	1734/8591	0.540 [0.450, 0.640]	**0.710 [0.620, 0.790]**
Qwen-7B	1808/8591	**0.620 [0.530, 0.710]**	0.610 [0.520, 0.700]
Mistral-7B	2079/8591	**0.640 [0.540, 0.730]**	0.450 [0.350, 0.550]
Mixtral-8x7B	1798/8591	**0.720 [0.630, 0.810]**	0.610 [0.510, 0.700]
LLaMA-70B	1222/8591	0.570 [0.470, 0.670]	**0.650 [0.560, 0.740]**
Qwen-72B	1043/8591	0.880 [0.810, 0.940]	**0.910 [0.850, 0.960]**

G LLM as a judge

Table 12: Literal → Metaphor generation flip rate by steering scale α and method (MM = mass-mean, DAS). Mean [95% CI]. Degenerate outputs counted as failed flips, few-shot prompts excluded. Only pairs where the judge gets the baseline explanation correct are considered (e.g. the meaning of "warm" in "the coffee is warm" is a model explanation such that the judge classifies it as "literal").

α	Method	LLaMA-8B	LLaMA-70B	Gemma-9B	Mistral-7B	Mixtral-8x7B	Qwen-7B	Qwen-72B
1	MM	0.05 [0.00, 0.10]	0.03 [0.00, 0.08]	0.02 [0.00, 0.05]	0.03 [0.00, 0.08]	0.05 [0.00, 0.11]	0.03 [0.00, 0.09]	0.08 [0.02, 0.15]
	DAS	0.02 [0.00, 0.05]	0.03 [0.00, 0.08]	0.00 [0.00, 0.00]	0.05 [0.00, 0.11]	0.02 [0.00, 0.05]	0.03 [0.00, 0.09]	0.02 [0.00, 0.05]
2	MM	0.12 [0.05, 0.20]	0.05 [0.00, 0.11]	0.11 [0.05, 0.18]	0.11 [0.03, 0.19]	0.11 [0.03, 0.19]	0.12 [0.05, 0.22]	0.09 [0.03, 0.17]
	DAS	0.03 [0.00, 0.08]	0.08 [0.02, 0.16]	0.03 [0.00, 0.08]	0.20 [0.11, 0.31]	0.03 [0.00, 0.08]	0.09 [0.02, 0.16]	0.05 [0.00, 0.11]
5	MM	0.44 [0.32, 0.56]	0.34 [0.23, 0.45]	0.33 [0.23, 0.44]	0.52 [0.39, 0.64]	0.38 [0.27, 0.50]	0.26 [0.16, 0.38]	0.29 [0.20, 0.40]
	DAS	0.34 [0.22, 0.46]	0.28 [0.17, 0.41]	0.12 [0.06, 0.21]	0.19 [0.09, 0.28]	0.16 [0.08, 0.25]	0.47 [0.34, 0.59]	0.15 [0.08, 0.25]
10	MM	0.20 [0.12, 0.31]	0.47 [0.36, 0.59]	0.55 [0.42, 0.67]	0.34 [0.23, 0.45]	0.47 [0.34, 0.59]	0.57 [0.45, 0.69]	0.11 [0.05, 0.18]
	DAS	0.25 [0.15, 0.37]	0.00 [0.00, 0.00]	0.03 [0.00, 0.08]	0.02 [0.00, 0.05]	0.06 [0.02, 0.12]	0.40 [0.28, 0.52]	0.43 [0.31, 0.55]

Table 13: Metaphor → Literal generation flip rate by steering scale α and method (MM = mass-mean, DAS). Mean [95% CI]. Degenerate outputs counted as failed flips; few-shot prompts excluded, only baseline-correct items.

α	Method	LLaMA-8B	LLaMA-70B	Gemma-9B	Mistral-7B	Mixtral-8x7B	Qwen-7B	Qwen-72B
1	MM	0.03 [0.00, 0.07]	0.01 [0.00, 0.04]	0.05 [0.01, 0.11]	0.06 [0.01, 0.11]	0.01 [0.00, 0.04]	0.04 [0.00, 0.09]	0.00 [0.00, 0.00]
	DAS	0.00 [0.00, 0.00]	0.01 [0.00, 0.04]	0.05 [0.01, 0.11]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.04 [0.00, 0.09]	0.00 [0.00, 0.00]
2	MM	0.04 [0.00, 0.09]	0.04 [0.00, 0.09]	0.11 [0.04, 0.18]	0.04 [0.00, 0.10]	0.07 [0.01, 0.14]	0.08 [0.03, 0.14]	0.00 [0.00, 0.00]
	DAS	0.01 [0.00, 0.04]	0.05 [0.01, 0.11]	0.08 [0.03, 0.14]	0.00 [0.00, 0.00]	0.01 [0.00, 0.04]	0.07 [0.01, 0.13]	0.03 [0.00, 0.07]
5	MM	0.03 [0.00, 0.07]	0.15 [0.07, 0.23]	0.25 [0.16, 0.36]	0.28 [0.18, 0.39]	0.24 [0.14, 0.34]	0.21 [0.12, 0.30]	0.08 [0.03, 0.15]
	DAS	0.09 [0.04, 0.16]	0.16 [0.08, 0.25]	0.12 [0.05, 0.20]	0.10 [0.04, 0.17]	0.11 [0.04, 0.19]	0.12 [0.05, 0.18]	0.19 [0.11, 0.27]
10	MM	0.03 [0.00, 0.07]	0.00 [0.00, 0.00]	0.13 [0.07, 0.21]	0.35 [0.24, 0.46]	0.33 [0.21, 0.44]	0.01 [0.00, 0.04]	0.03 [0.00, 0.07]
	DAS	0.00 [0.00, 0.00]	0.05 [0.01, 0.11]	0.03 [0.00, 0.07]	0.06 [0.01, 0.11]	0.24 [0.14, 0.34]	0.05 [0.01, 0.11]	0.25 [0.16, 0.35]

H Example generations

	Output	Steered?
Metaphorical → Literal		
clear thinker (CLARITY)		
Baseline	easy to understand and free from confusion.	—
MM	easy to understand and free from confusion.	✗
DAS	easy to understand and not ambiguous.	✗
smooth player (TEXTURE)		
Baseline	having a calm and confident manner.	—
MM	having a calm and peaceful appearance.	✗
DAS	having a flat, even surface, and not necessarily emotionally complex.	✓
Literal → Metaphorical		
rough landscape (TEXTURE)		
Baseline	having a harsh, uneven surface.	—
MM	having an uneven surface with many obstacles.	✗
DAS	having an uneven surface with many obstacles.	✗
clear broth (CLARITY)		
Baseline	transparent and free from impurities.	—
MM	transparent and easy to understand.	✓
DAS	transparent and free from impurities.	✗

Table 14: Qualitative examples of steering outcomes for mass-mean (MM) and DAS interventions. ✓ indicates the judge's label flipped to the steered direction. ✗ indicates no flip.

I Matched-n control

Concept	Full n (global DAS)	Matched n (M_n, global DAS)	Domain-specific DAS (Dom)	M_n − Dom [95% CI]
clarity	0.581	0.513	0.441	$+0.073$ $[+0.003, +0.145]$
depth	0.924	0.855	0.227	$+0.628$ $[+0.598, +0.662]$
light	0.580	0.412	0.524	-0.113 $[-0.158, -0.068]$
strength	0.919	0.850	0.776	$+0.074$ $[+0.046, +0.100]$
substance	0.643	0.524	0.500	$+0.024$ $[-0.095, +0.143]$
taste	0.830	0.750	0.587	$+0.163$ $[+0.083, +0.239]$
temperature	0.650	0.447	0.583	-0.135 $[-0.187, -0.084]$
texture	0.767	0.587	0.506	$+0.080$ $[+0.030, +0.132]$

Table 15: We retrain the global DAS matrix on a number of samples for each domain, matching the smallest domain-specific training domain ($n = 255$). For 6/8 cells, we achieve a comparable transfer IIA score to the global DAS matrix where no such filter is applied. Results reported for Gemma 2 9B.

J Steering meaning

J.1 NIE calculation

For *identification* task, we first compute PD^+ and PD^- as the average probability differences $P(\text{metaphoric}) - P(\text{literal})$ over examples from the metaphoric$(+)$ and literal$(-)$ classes, respectively. After intervention, we obtain PD_*^+ and PD_*^- in the same way. The NIE is then computed as:

$$\text{NIE}_{\text{lit}\to\text{met}} = \frac{PD_*^- - PD^-}{PD^+ - PD^-}$$

$$\text{NIE}_{\text{met}\to\text{lit}} = \frac{PD_*^+ - PD^+}{PD^- - PD^+}.$$

Therefore, an NIE of 0 indicates no causal effect, while an NIE of 1 corresponds to a complete shift in model predictions.

J.2 Full NIE results

train/test	probe	LLaMA-8B		LLaMA-70B		Gemma-9B		Gemma-27B	
		inapt→apt	apt→inapt	inapt→apt	apt→inapt	inapt→apt	apt→inapt	inapt→apt	apt→inapt
Adj/Adj	LogReg	.038	.051	.033	.028	**.176**	**.211**	**.293**	**.774**
	MM-diff	**.227**	**.471**	**.183**	**.168**	.149	.097	.262	.387
Noun/Noun	LogReg	.025	.040	.029	.029	**.220**	.060	**.227**	.282
	MM-diff	**.195**	**.355**	**.216**	**.126**	.212	**.110**	.045	**.304**
Adj/Noun	LogReg	.026	.032	.029	.031	**.211**	**.297**	**.076**	**.763**
	MM-diff	**.138**	**.254**	**.189**	**.124**	.177	.223	.050	.194
Noun/Adj	LogReg	.028	.040	.019	.015	.076	.072	**.449**	.429
	MM-diff	**.230**	**.666**	**.121**	**.171**	**.110**	**.131**	.230	**.497**

train/test	probe	Mistral-7B		Mixtral-8x7B		Qwen-7B		Qwen-72B	
		inapt→apt	apt→inapt	inapt→apt	apt→inapt	inapt→apt	apt→inapt	inapt→apt	apt→inapt
Adj/Adj	LogReg	.014	.095	.075	.178	.035	.054	.100	.084
	MM-diff	**.075**	**.449**	**.093**	**.292**	**.676**	**.578**	**.251**	**.503**
Noun/Noun	LogReg	.007	**.033**	.013	.042	.010	.031	.097	**.072**
	MM-diff	**.023**	.025	**.031**	**.051**	**.467**	**.081**	**.243**	.065
Adj/Noun	LogReg	.006	.030	.012	.123	.050	.056	.097	.040
	MM-diff	**.024**	**.153**	**.015**	**.159**	**.397**	**.561**	**.167**	**.397**
Noun/Adj	LogReg	.006	.056	.042	.123	.038	.031	.072	.060
	MM-diff	**.054**	**.520**	**.093**	**.367**	**.553**	**.353**	**.285**	**.318**

Table 16: Normalized indirect effects (NIEs) of mass-mean difference (MM-diff) and logistic regression (LogReg) probes on the *identification* task for eight models. Higher NIE values indicate stronger causal steering effects. Higher scores are highlighted in bold.

train/test	probe	LLaMA-8B		LLaMA-70B		Gemma-9B		Gemma-27B	
		inapt→apt	apt→inapt	inapt→apt	apt→inapt	inapt→apt	apt→inapt	inapt→apt	apt→inapt
Noun	LogReg	.011	.014	.001	.005	.009	.031	.038	.144
	MM-diff	**.355**	**.564**	**.126**	**.305**	**.155**	**.353**	**.304**	**.656**

train/test	probe	Mistral-7B		Mixtral-8x7B		Qwen-7B		Qwen-72B	
		inapt→apt	apt→inapt	inapt→apt	apt→inapt	inapt→apt	apt→inapt	inapt→apt	apt→inapt
Noun	LogReg	.003	.013	.003	.017	.002	.008	.007	.033
	MM-diff	**.025**	**.165**	**.051**	**.238**	**.081**	**.261**	**.065**	**.344**

Table 17: Normalized indirect effects (NIEs) of mass-mean difference (MM-diff) and logistic regression (LogReg) probes on the *interpretation* task for eight models. Probes are trained and tested on our Noun dataset. Higher NIE values indicate stronger causal steering effects. Higher scores are highlighted in bold.

J.3 Flip rate results

Task	Model	Data	Layer	Original label	Before	After	Flip rate
T1	LLaMA-3.1-8B	Noun	11	metaphor	27	13	51.85%
	LLaMA-3.1-8B	Noun	11	literal	28	7	75.00%
	LLaMA-3.1-8B	Adj	11	metaphor	21	14	33.33%
	LLaMA-3.1-8B	Adj	11	literal	7	2	71.43%
	Gemma-2-27B	Noun	20	literal	23	18	21.74%
	Gemma-2-27B	Adj	20	literal	17	12	29.41%
	Qwen2.5-72B	Noun	54	literal	26	18	30.77%
	Qwen2.5-72B	Adj	54	literal	18	8	55.56%
	Mixtral-8x7B	Noun	14	literal	20	14	30.00%
	Mixtral-8x7B	Adj	14	literal	13	10	23.08%
	LLaMA-3.1-70B	Noun	29	literal	25	22	8.00%
	LLaMA-3.1-70B	Adj	29	literal	18	16	11.11%
T2	Gemma-2-27B	Noun	20	inapt	26	21	19.23%
	Gemma-2-9B	Noun	20	inapt	26	23	11.54%
	Qwen2.5-72B	Noun	54	inapt	25	24	4.00%
	Mixtral-8x7B	Noun	14	inapt	27	25	7.41%

Table 18: Steering results using mass-mean probes on the identification (T1) and interpretation (T2) tasks. We intervene on correctly classified test examples and steer them toward the opposite label. "Before" and "After" indicate the numbers of examples retaining the original label before and after steering, respectively. Flip rate denotes the proportion of originally correct predictions that are flipped to the opposite label.

J.4 Prompt format

All models use the same plain-continuation meaning-format prompt with five hardcoded few-shot examples demonstrating free-form meaning descriptions (no classification labels). The test query asks for the meaning of the adjective in context:

The welcome was warm. The meaning of 'warm' is: affectionate and emotionally inviting.
The lamp was bright. The meaning of 'bright' is: emitting a strong, clear light.
The novel was deep. The meaning of 'deep' is: intellectually complex and rich in meaning.
The stone was smooth. The meaning of 'smooth' is: having an even surface with no roughness.
The defeat was bitter. The meaning of 'bitter' is: painful, hard to accept, and leaving resentment.
The [noun] was [adj]. The meaning of '[adj]' is:

No chat template is applied to any model. The hook patches the hidden state at the last token position at every forward step during generation. The token position is the same token position used for DAS training.

K Prompt for LLM judge

You are classifying whether a meaning description conveys a literal or metaphorical sense of an adjective.

A LITERAL meaning describes a physical, concrete, or sensory property (e.g. temperature, texture, weight, brightness).
A METAPHORICAL meaning describes an abstract, figurative, or emotional property (e.g. mood, personality, intellectual quality).

Reply with exactly one word: either "literal" or "metaphorical".
No explanation. Do not repeat the sentence.
Sentence: "The noun was adj."
Adjective: 'adj'
Meaning description: "first sentence of completion"
Is this meaning literal or metaphorical?

L Dataset Usage

We use datasets from Liu et al. (2022b); Gutiérrez et al. (2016); Tsvetkov et al. (2014). While we slight alter the format of these datasets, all are used for their original purpose, namely metaphor research. All datasets are freely available in the public domain.

M Computational resources

All experiments took a total of about 125 hours on an H100 GPU. Hyperparameters checked include steering strength $\alpha \in \{1, 2, 5, 10\}$ for generation experiments and subspace dimension $k \in \{4, 8, 16, 32\}$ for DAS matrix training.

N Use of AI assistants

AI assistants are used for code debugging, generating the prompt for the generation experiment, and generating tables. Numbers are double-checked manually.