BEYOND LINEARITY: THE CONTEXTUAL GEOMETRY OF CONCEPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representations of concepts are commonly read as directions within neural activations, assuming that a concept is mediated by a single, context-free vector. We prove that this assumption fails in general because modern architectures inherently encode contextual interactions. Across fifteen models and three modalities, the directions corresponding to the same concept exhibit high variability across contexts. Furthermore, measuring their effective dimension reveals that concepts correspond to multidimensional structures in the activation space. This motivates our novel contextual representations of concepts, which are able to capture these structures across different contexts. Under causal intervention on LLMs, or steering, they allow consistent reductions in perplexity over single directions, highlighting more faithful representation of concepts, and improved control over the behavior of models.

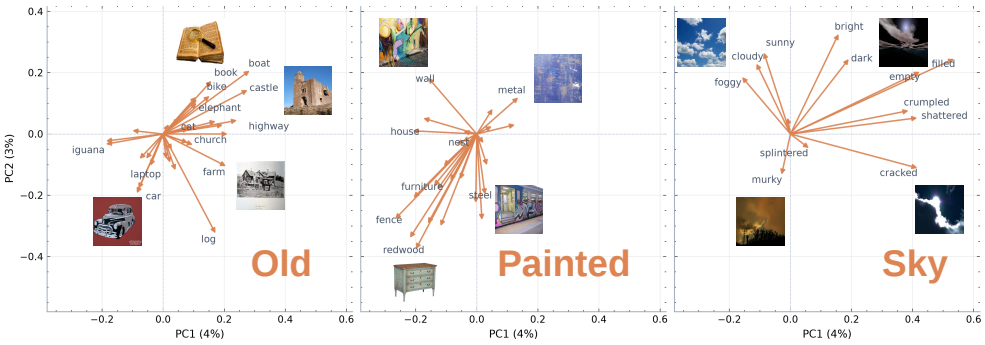

Figure 1: Contextual representations of the attributes “Old”, “Painted” and “Sky” in CLIP activations (MIT States), projected on the two first principal components. The same attributes point highly different directions across contexts (written at the tip of corresponding arrows).

1 INTRODUCTION

Extracting representations of interpretable concepts within neural networks activations is crucial for understanding deep learning systems. However, the definition of a *concept* is non-trivial, and has been subject to debates in machine learning research (Koh et al., 2020; Bricken et al., 2023; Park et al., 2024; Sharkey et al., 2025), but also in social sciences such as philosophy (Hegel, 1816; Deleuze, 1968), linguistics (de Saussure et al., 1916; Lakoff, 1987) or cognitive sciences (Rosch, 1978; Medin & Schaffer, 1978). In its simplest possible form, a concept is a *semantic unit*: debates concern their origin and structure. Our objective is to study concepts within neural activations, in order to assess, refine, and propose alternatives to popular hypotheses in mechanistic interpretability.

Mechanistic interpretability research seeks interpretable concepts in the activations of neural networks, with the objective of understanding and control their behavior. The majority of the field focuses on concepts represented as single directions in the activation space (Elhage et al., 2022; Park

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2024). Works aim at studying specific behaviors of neural networks such as refusal (Arditi et al., 2024) and harmfulness (Zhao et al., 2025). Unsupervised techniques have received massive interest, in particular sparse autoencoders (Bricken et al., 2023), which address the question "What concepts do models represent?" by learning a sparse decomposition from the activations. They have been applied to a wide range of models, spanning multiple modalities (Thasarathan et al., 2025; Miao et al., 2025; Mariotte et al., 2026). Steering methods (Turner et al., 2025; Zhou et al., 2025; Ahsan & Wallace, 2026) use linear representations of concepts to intervene on the downstream behavior of models.

In order to extend the understanding of neural networks, some works also focus on the analysis of specific concepts and mechanisms, to go beyond linearity. As different concepts can be semantically related, multiple works study how language models represent hierarchies (Park et al., 2025; Costa et al., 2025). Studying specific concepts reveal that some concepts such as "days of the week" may form circular structures activations (Engels et al., 2025). Arithmetic operations have been shown to respect a similar behavior (Feucht et al., 2026), highlighting another case where concepts may not be properly represented as single directions.

In order to address the limitations of current representations of concept, and propose a solution, our main contributions are:

- A theoretical analysis of how modern architectures break the linearity of concepts. We prove that they fail to represent concepts as linear if they compute with them.
- An empirical characterization of the variability and size of concepts across fifteen different models, nine datasets and three modalities. Concept representations thus appear highly context-dependent.
- CACTUS: a novel steering method that uses this contextual information to better control the behavior of language models. CACTUS is able to consistently reduce perplexity during concept insertion.

2 Related Works

Linear representations of concepts The major line of research in interpretability research considers the concepts represented by neural networks to be linear in their activations, with roots in early word embedding techniques (Mikolov et al., 2013). In particular, probe-based concept representation has been extensively studied in both NLP (Gupta et al., 2015), computer vision (Alain & Bengio, 2017) and reinforcement learning (Lovering et al., 2022). Contrast-Consistent Search probes for an axis in the activation space, using contrastive groups corresponding to presence or absence of predefined concepts (Burns et al., 2023). TCAV (Kim et al., 2018) and ACE (Ghorbani et al., 2019) represent concepts using a predefined list. Works suggest that refusal (Arditi et al., 2024) and harmfulness (Zhao et al., 2025) may each correspond to a single axis. In a similar fashion, DiffMean (Marks & Tegmark, 2024) represents the truth or falsehood of statements linearly in the activation space. Park et al. (2024) formalizes this assumption as *The Linear Representation Hypothesis*, and studies its applications to LLM unembeddings.

Unsupervised concept discovery Several methods represent concepts as linear, and extract them without supervision to understand *what* concepts do neural networks represent? Most of those methods are sparse autoencoders (SAE) (Bricken et al., 2023), that compute a sparse decomposition of activations, motivated by the superposition hypothesis (Elhage et al., 2022). Multiple extensions have been proposed to enhance the quality of learned concepts (Rajamanoharan et al., 2024; Gao et al., 2025; Bussmann et al., 2024), or their consistency (Fel et al., 2025). SAE have been massively used to study textual representation, but also image encoders (Thasarathan et al., 2025), audio (Mariotte et al., 2026) and 3D representations (Miao et al., 2025). Deleuzian concepts are an alternative method that models differences between activations (Cornet et al., 2026).

Structured concepts While most works consider that concepts are linear and independent from one another, some try to go beyond this representation, by representing some structure in specific cases. In particular, works study atomicity of SAE concepts (Leask et al., 2025; Levinson, 2026). Similarly, hierarchical structures have been analyzed with SAE (Zaigrajew et al., 2025; Costa et al.,

2025) and concept bottleneck models (Hill et al., 2026), and have been shown to form orthogonal structures (Park et al., 2025). Some works propose to nuance the linear representation hypothesis for recurrent networks (Csordás et al., 2024) or LLM refusal (Joad et al., 2026), suggesting different geometric structures in neural activations. A recent line of work models some specific concepts as sequential (Mishra et al., 2026), or even circular structures (Engels et al., 2025; Feucht et al., 2026). Unsupervised methods have been used to compare multiple models, in order to analyze their differences at the concept-level (Wang et al., 2025; Cornet et al., 2025; Stevens et al., 2025).

Steering Understanding how a model represents some concept allows modifying the neural activations towards this concept: this operation is called *Steering* (Turner et al., 2025). Most steering methods represent concepts linearly, and steering therefore consists in the addition of a fixed vector (Zou et al., 2023; Li et al., 2023; Marks & Tegmark, 2024; Turner et al., 2025). Other approaches rely on fine-tuning (Hu et al., 2022; Wu et al., 2024; 2025; Sun et al., 2025), and optimize steering directions on a given dataset in task-specific settings. Finally, unsupervised methods such as sparse autoencoders have been used for steering multimodal representations (Pach et al., 2025) and for bias and hallucination reduction in language models (Zhou et al., 2025; Parekh et al., 2025). Ahsan & Wallace (2026) use steering to analyze sensitive biases in foundation models. Recent works propose to study the correspondence between behavior and representation in the context of steering (Mishra et al., 2026; Feucht et al., 2026; Wurgaft et al., 2026).

3 A THEORETICAL STANDPOINT ON THE GEOMETRY OF CONCEPTS

3.1 LINEAR REPRESENTATIONS OF CONCEPTS

The objective of this study is to analyze inner concepts in neural activations, and determine whether they may or may not be considered as globally linear. Let $X \in \mathbb{R}^d$ be the d-dimensional activations of a model M at layer l for a dataset $\mathcal{D}$.

> **Definition 1. Concept:** A concept is a function $\delta : \mathbb{R}^d \to \mathbb{R}$ that maps activations to a scalar score representing the presence or intensity of a semantic attribute.

The most standard view of concepts is to represent them as globally linear among neural activations (Bricken et al., 2023; Burns et al., 2023; Park et al., 2024; Cornet et al., 2026). In other words, it amounts to representing concepts as vectors in the activation space. Therefore, for the concept c

$$\exists \mathbf{v}_c \in \mathbb{R}^d, \forall \mathbf{x} \in \mathbb{R}^d \quad \delta_c(\mathbf{x}) = \mathbf{v}_c \cdot \mathbf{x} \tag{1}$$

In our later experiments, we model linear concepts corresponding to categorical attributes $C = c$ of the dataset. We represent a linear concept as a discriminant between activations of X upon samples corresponding to the targeted attribute and the others. Then, we represent:

$$\mathbf{v}_c = \mathbb{E}[X|C = c] - \mathbb{E}[X|C \neq c] \tag{2}$$

That is equivalent to the DiffMean method proposed by Marks & Tegmark (2024). Furthermore, it is worth noticing that concept representation methods based on a linear representation (sparse autoencoders, DiffMean, CCS...) all consider concepts as *globally* linear, meaning that $\mathbf{v}_c$ is constant across $\mathbb{R}^d$. Note that those hypotheses must hold at *every* layer, as interpretability techniques using them are applied at multiple different stages of models (Gao et al., 2025; Wu et al., 2025).

3.2 TRANSFORMERS BREAK LINEARITY

The majority of modern foundation models are based on the transformer architecture (Vaswani et al., 2017). Therefore, analysing this specific architecture may allow us to uncover mechanistic insight on the concepts transformer-based models represent. The question we address here is: *can a concept inserted along a fixed direction stay linear through a transformer block?*

Proposition 1: Let $B = \phi \circ \mathrm{Attn}$ be a transformer block on $\mathbb{R}^d$, comprising an attention module Attn with H heads, and a position-wise mapping ϕ (MLP, normalization layers...). Let c be a concept that is linearly represented along a vector v at the input of B. If c is linear at the output of B, then the attention module does not compute with c. At most d/H such concepts may be carried per head.

Implications Proposition 1 has severe implications on the geometry of concepts. Only specific vectors v can respect conditions for linearity through a transformer block. A concept that is linear at some layer stays linear at the next only if attention does not compute with it: considering concepts as linear by default fails for concepts a model actually use. At most d/H concepts can be carried linearly per head through attention, that highly limits the number of linear concepts that should be considered simultaneously in a model. This should discourage representations of overcomplete set of linear concepts: sparse autoencoders. Those conditions apply on both model weights and concept direction v. Most concept-based methods extract concepts from neural activations, not weights, and in practice do not extract concepts matching those specific conditions.

Proof Let $x = (x_1, ..., x_n)$ be the activations of n tokens, $x_i \in \mathbb{R}^d$, entering a transformer block B, that is composed of an attention module Attn, and ϕ a sequence of position-wise mappings (feed-forward network, possibly normalization layers), such that $B = \phi \circ \mathrm{Attn}$. consider $\tilde{x}$, that only differs from x because of one linear concept being present at position s. With $v \in \mathbb{R}^d, \varepsilon \in \mathbb{R}^*$, we obtain at position i: $\tilde{x}_i = x_i + \varepsilon v \cdot \mathbb{1}[i = s]$. This concept being linear at the output of B requires existence of $v_{out} \in \mathbb{R}^d, B(\tilde{x})_i - B(x)_i = \gamma v_{out}, \forall x$, with γ possibly depending on x, i, s and ε, while the output direction v_{out} remains fixed for every input sequence

The position-wise mapping ϕ receives only the activation $\mathrm{Attn}(\tilde{x})_i$ at one position. Different sequences can have the same attention output at some position, yet inserting a concept v moves it in different directions depending on different contexts. For ϕ to restore linearity would need it to send all these displacements onto a single direction: v_{out}. ϕ would then erase all information that was computed through the attention except for v_{out}. Therefore, linearity at the output of B requires linearity at the output of the attention, so that

$$\exists v_{attn} \in \mathbb{R}^d, \forall x, \mathrm{Attn}(\tilde{x})_i - \mathrm{Attn}(x)_i = \gamma v_{attn} \tag{3}$$

Formalizing Attn as a multi-head attention, with H different heads gives

$$\mathrm{Attn}(x)_i \;=\; x_i \;+\; \sum_{h=1}^{H} \sum_{j=1}^{n} A^h_{i,j}\, W^{OV}_h\, x_j, \qquad A^h_{i,j} \;=\; \mathrm{softmax}_j\left(\frac{x_i^\top W^{QK}_h\, x_j}{\sqrt{d/H}}\right) \tag{4}$$

with i the position the model reads at, j the position being read from, $W^{QK}_h = (W^Q_h)^\top W^K_h$ and $W^{OV}_h = W^O_h W^V_h$ the QK (Query-Key) and OV (Output-Value) circuits of head h (Elhage et al., 2021). Computing the difference at the output of attention gives

$$\mathrm{Attn}(\tilde{x})_i - \mathrm{Attn}(x)_i = \underbrace{\varepsilon v\, \mathbb{1}[i = s]}_{\text{residual connection}} + \underbrace{\varepsilon \sum_h \tilde{A}^h_{i,s}\, W^{OV}_h v}_{\text{value read from } s} + \underbrace{\sum_h \sum_j (\tilde{A}^h_{i,j} - A^h_{i,j})\, W^{OV}_h x_j}_{\text{change of attention}} \tag{5}$$

The change of attention in eq. (5) if built from every token activation x_j, varying freely across sequences. Therefore, it represents no fixed direction, and to allow the total difference in eq. (5) to be linear, this term must collapse to zero. In particular, to do so for every x, the attention must specifically give $\tilde{A}^h_{i,j} = A^h_{i,j}, \forall i, j, h$. At position $i \neq s$, the concept insertion moves one attention logit by $\varepsilon x_i^\top W^{QK}_h v$. However at position s, every logit is moved by $\varepsilon v^\top W^{QK}_h x_j$. Therefore, to be verified for all x, eq. (3) forces

$$W^K_h v = W^Q_h v = 0, \forall h \tag{6}$$

Considering that the change of attention as null, linearity at the output of the attention still requires the sum of the residual connection and the value read from s to be linear. As the residual connection forces v at position s, it constrains v_{attn} to be colinear with v. Therefore, the directions $W_h^{OV} v$ read from s must be colinear with v, hence v must be an eigenvector of every OV circuit. Only d/H such concepts may then exist per head

$$\forall h, \exists \lambda_h \in \mathbb{R}, W_h^{OV} v = \lambda_h v \tag{7}$$

An inserted concept v respecting both eqs. (6) and (7) would then exit the attention as

$$\text{Attn}(\tilde{x})_i - \text{Attn}(x)_i = \varepsilon v \left(\mathbb{1}[i = s] + \sum_h A_{i,s}^h \lambda_h \right) \tag{8}$$

Therefore, linearity of v through a transformer block B requires v to only be rescaled during the attention, never used to route nor transformed. Linearity through a transformer block thus fails for every concept v the attention computes with.

4 Empirical characterization: are contexts globally linear?

After theoretically showing that concepts a transformer actually computes with cannot be linear through attention, we characterize this property empirically. Our experiments comprise fifteen different models and nine datasets, spanning image, text and audio modalities, with multiple different architectures and training paradigms. More details are reported in section E.

Image. We consider both transformers and convolutional architectures, namely DinoV3 (Siméoni et al., 2025), CLIP (Radford et al., 2021), SwinV2 (Liu et al., 2022a), ResNet50 (He et al., 2016) and ConvNext (Liu et al., 2022b). They are studied on the art dataset WikiArt (Tan et al., 2018), DomainNet (Peng et al., 2019) that is designed for domain adaptation, and MIT States (Isola et al., 2015) that is built for compositional learning.

Text. We analyze 2 encoder models, XLM-Roberta (Conneau et al., 2020) and mDeBERTaV3 (He et al., 2021), as well as 3 decoders : Gemma3 (Team, 2026), Llama3.1 (Grattafiori et al., 2024) and Qwen3.5 (Qwen Team, 2026). Regarding datasets, GoEmotions (Demszky et al., 2020) is formed of internet messages, while MTOP (Li et al., 2021) and MASSIVE (FitzGerald et al., 2022) are multilingual datasets for semantic parsing and natural language understanding.

Audio. We consider both speech only and general-purpose audio encoders, namely HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), Whisper (Radford et al., 2023), CLAP (Elizalde et al., 2023) and an audio-spectrogram transformer (Gong et al., 2021). We analyze them on the emotion recognition datasets ESD (Zhou et al., 2021) and CREMA-D (Cao et al., 2014), as well as AudioMNIST (Becker et al., 2024).

4.1 Diversity of concepts across contexts

From section 3.2, concepts a transformer model computes with cannot be globally linear. Then, one may question how this theoretical result translates empirically. We introduce contextual concepts, in order to study how concepts are globally linear or not.

> **Definition 2. Contextual concepts:** Concepts extracted within a specific context, with samples sharing similar semantics (i.e. dataset labels). The contextual representation of concept c in context k is $\delta_c^{(k)}(\mathbf{x}) = \mathbf{v}_c^{(k)} \cdot \mathbf{x}$, with
>
> $$\mathbf{v}_c^{(k)} = \mathbb{E}[X|C = c, K = k] - \mathbb{E}[X|C \neq c, K = k]$$

If concepts are globally linear across contexts, then $\forall c, k, k', \cos(\mathbf{v}_c^{(k)}, \mathbf{v}_c^{(k')}) = 1$.

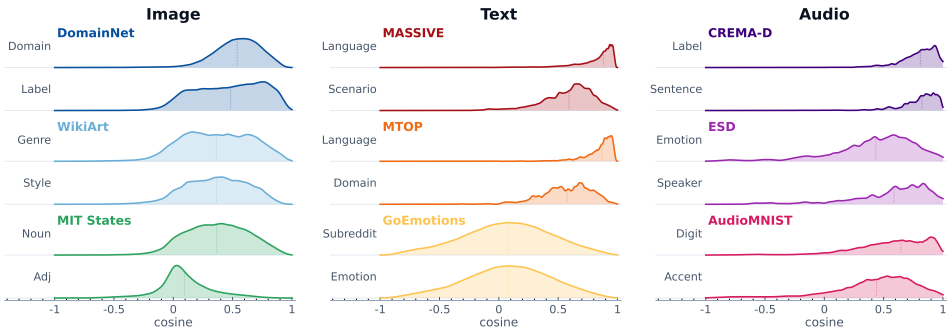

Figure 2: Pairwise cosines of contextual directions $\mathbf{v}_a^b$ across contexts, aggregated across models.

We study how different contexts $K = k$ can lead to variability, and diversity across contextual concepts for a same attribute c. We measure the distribution D of contextual concepts, per attribute c

$$D = \left\{ \cos(\mathbf{v}_c^{(k)}, \mathbf{v}_c^{(k')}) | k \neq k', (C = c \cap K = k) \neq \emptyset \right\} \tag{9}$$

Figure 1 highlights high variety for an individual concept across contexts, highlighting that the $\mathbf{v}_c^{(k)}$ are not restricted to a single direction, or even to a cone. Figure 2 shows aggregated results across models, for each dataset attribute, at models' last layer. For example, with MIT States, it shows the variability of "adj" representations across "noun", and respectively the opposite. It empirically demonstrates a high variability with most cosines being far below 1 (that would match the linear hypothesis), with a high proportion of similarities even being negative. Image models seem to encounter the highest variability across all three modalities. For text, GoEmotions, that is formed of short messages from different online communities has a very high variability for both attributes. Language has, for both MTOP and MASSIVE the highest median pairwise cosine, meaning that it is the closest to be linear in practice at the last layer of models.

Section A.1 contains detailed and numerical results across models, including results at intermediate layers of studied models. Note that, as our analysis comprises some non-transformer models, empirical results show that they are not the only architecture to break linearity.

4.2 Size of concepts

Once we have shown that concepts are not *globally linear*, we might want to characterize *how*: what is their size? For each attribute of dataset $\mathcal{D}$ and a model M at layer l, we compute G the Gram matrix of its corresponding contextual concepts, with

$$G_{ij} = \cos\left(\mathbf{v}_c^{(i)}, \mathbf{v}_c^{(j)}\right) \tag{10}$$

This Gram matrix models the similarity structure induced by c within the activations of M. Our objective is then to determine its size, its effective dimension d_{eff} (Recanatesi et al., 2022). With λ_i the eigenvalues of G_{ij}

$$d_{\text{eff}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \tag{11}$$

The values of d_{eff} are bound between one, and the number of contexts the studied context appears in (case of global independence, if the number of contexts is less than d). They indicate an estimation of the *true dimensionality* of the concept c.

From table 1, concepts are represented by multi-dimensional structures. Those results show that image concepts not only have the highest variability, but seem to occupy the largest number of

Table 1: Effective dimension of concepts, mean ± std. Results aggregated between models. "Contexts" indicates the average number of contexts each concept is present in.

Modality	Dataset	Concept	Contexts	d_{eff} (intermediate)	d_{eff} (last)
Image	DomainNet	domain	345.0	1.99 ± 0.54	3.65 ± 1.10
	DomainNet	label	6.0	3.08 ± 0.48	2.50 ± 0.38
	WikiArt	genre	22.2	4.85 ± 0.60	5.01 ± 0.87
	WikiArt	style	9.0	3.09 ± 0.26	3.80 ± 0.79
	MIT States	noun	8.0	3.51 ± 0.13	3.67 ± 0.21
	MIT States	adj	17.4	7.44 ± 1.46	9.96 ± 2.27
Text	MASSIVE	language	18.0	1.64 ± 0.36	1.31 ± 0.35
	MASSIVE	scenario	7.0	2.00 ± 0.98	2.21 ± 0.36
	MTOP	language	11.0	1.68 ± 0.36	1.30 ± 0.16
	MTOP	domain	5.0	1.94 ± 0.80	2.12 ± 0.27
	GoEmotions	subreddit	20.1	10.00 ± 5.40	6.84 ± 2.07
	GoEmotions	emotion	346.9	24.19 ± 19.74	10.36 ± 4.63
Audio	CREMA-D	label	12.0	1.38 ± 0.07	1.50 ± 0.20
	CREMA-D	sentence	6.0	1.36 ± 0.13	1.38 ± 0.20
	ESD	emotion	10.0	2.86 ± 0.55	2.97 ± 0.44
	ESD	speaker	5.0	1.91 ± 0.27	1.94 ± 0.25
	AudioMNIST	digit	17.0	2.05 ± 0.52	2.27 ± 0.99
	AudioMNIST	accent	10.0	3.05 ± 1.10	3.09 ± 1.04

dimensions. Also, comparing d_{eff} at the last layer and at intermediate layers, there is no clear tendency of "inflation" nor "deflation" throughout the model. Concepts extracted from GoEmotions (online messages) obtain very large d_{eff} in every case, corresponding to highly multidimentional concepts. The domain in DomainNet is an attribute that appears in 345 contexts on average, but this information gets highly compressed at less than 2 dimensions at the intermediate layers. Per-model results are provided in section A.3.

5 CACTUS: Context-Aware ConcepT Union Steering

Steering is a popular practice in mechanistic interpretability, that consists in direct intervention on a model's internal representations, to modify or control its behavior upon a chosen concept (Turner et al., 2025). The objective is to use a learned concept representation to perform semantic editing within neural activations. Steering is usually used on intermediate representations, therefore our main experiments consider the middle layer of each model (e.g. 21st layer for a model with 42). Complementary experiments are presented on the last layer of models in section B.

Linear representations of concepts allow steering, by computing the steered activation as $x + \alpha \mathbf{v}_c$, therefore simply moving along $\mathbf{v}_c$ with magnitude α (normalized displacement length).

A trivial non-linear representation is to use a quadratic classifier (Quadratic Discriminant Analysis). We perform iterative gradient ascent along the classifier's gradient, to effectively insert a concept.

As we demonstrated theoretically and empirically that concepts vary upon context, we may use this information for representing and steering concepts. We consider a concept as strongly activated as any of its corresponding contextual concepts, including those where c is the represented attribute, and those where it corresponds to the concept. Therefore, we introduce CACTUS (Context-Aware ConcepT Union Steering), that models concepts by considering the union of every contextual concepts representing the targeted attribute, or using it as context.

$$\delta_c(\mathrm{x}) = \max_k \max(\mathbf{v}_c^{(k)} \cdot \mathrm{x}, \mathbf{v}_k^{(c)} \cdot \mathrm{x}) \tag{12}$$

This representation is piecewise linear, but the combination amounts to a highly non-linear function. To illustrate this, Figure 3 represents for each method the concept value and gradient (or scoring

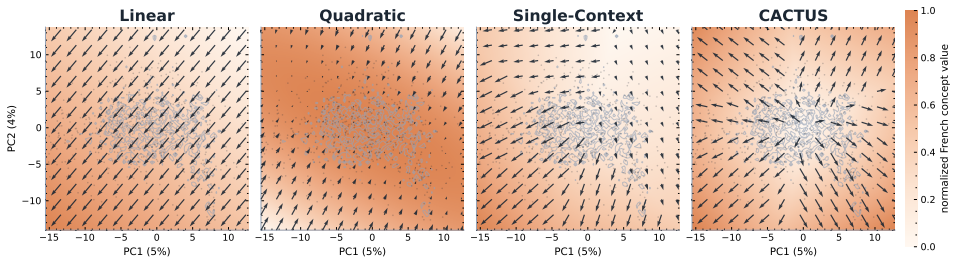

Figure 3: Projection of different representations of the concept "French" (MTOP language) in Qwen3.5-9B activations at layer 16

direction) for the "*French language*" concept from MTOP dataset in the activations of Qwen3.5-9B. Steering is done by moving the activation along the gradient of δ_c, being $\arg\max_k\{\mathbf{v}_c^{(k)}\cdot \mathrm{x}, \mathbf{v}_k^{(c)}\cdot \mathrm{x}\}$. For faithful comparison, we also evaluate "Single-context" steering, that is $\delta_c(\mathrm{x}) = \max_k(\mathbf{v}_c^{(k)}\cdot \mathrm{x})$, without considering the contextual representation of the attribute k with c as context.

We use a perplexity-based evaluation to assess the performances of the different steering methods on autoregressive text decoders. Perplexity is a measure of the "surprise" of a model, given a sequence of tokens. Our evaluation of steering consists in measuring this perplexity under steering: if we steer the concept of "joy" in neural activations, a LLM should get less surprised for a sequence that is actually joyful. Using dataset labels, we insert a concept corresponding to sequence-level semantics at every position with magnitude α, and measure the steered perplexity as

$$\mathrm{PP}_\alpha = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log p_\alpha(w_t \mid w_{<t})\right) \tag{13}$$

To obtain the optimal steering magnitude α, that is optimally able to insert concept information, we perform a sweep on α values. Figure 4 shows the geometric mean across sequences of $\mathrm{PP}_\alpha/\mathrm{PP}_0$, representing the mean perplexity reduction that different steering methods allow (the lower the better), for steering emotions (GoEmotions) and language (MTOP, MASSIVE). On all tested models, CACTUS reaches the lowest perplexity on average, by a 9% margin over linear steering. CACTUS is technically the combination of Single-Context steering, that is a strong baseline (fig. 4), and the contextual representations where the targeted concept is considered at the context. The latter is a non-effective method when considered in isolation, reaching a mean $\mathrm{PP}_\alpha/\mathrm{PP}_0 > 1$ on every task, but the union with Single-Context is what allows steering improvement. Furthermore, explicitely using dataset information regarding the context k (context-oracle) forms a lower-bound for CACTUS's perplexity. Linear steering is a reliable baseline upon this metric, as it is consitently able to obtain $\mathrm{PP}_\alpha/\mathrm{PP}_0 < 1$ across models and datasets. On the other end, quadratic steering underperforms the simple linear baseline on every model. This shows that non-linearity is not sufficient for efficient steering, modeling the diversity of contexts is what allows proper causal intervention in neural activations.

To be usable in real-world scenarios, a steering method must allow meaningful concept editing at different α values, so the steering magnitude choice does not cause hard failures. Figure 5 shows mean $\mathrm{PP}_\alpha/\mathrm{PP}_0$ swept upon α. From those results, steering performance is smooth upon α, and all the tested methods have similar ordering across steering magnitudes on average. Therefore, results from fig. 4 are robust over α, and CACTUS allows consistent steering improvement.

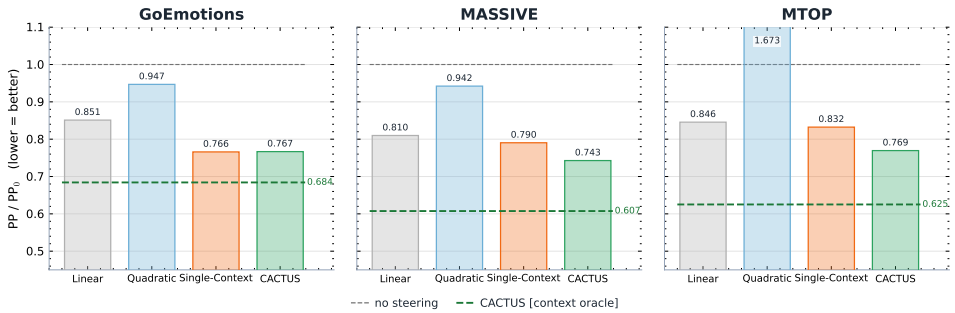

Figure 4: Perplexity reductions of different steering methods for GoEmotions emotion and language (MASSIVE, MTOP), intermediate layers. Results mean-aggregated across Qwen3.5-9B,Gemma4-E4B and Llama3.1-8B. The lower the better.

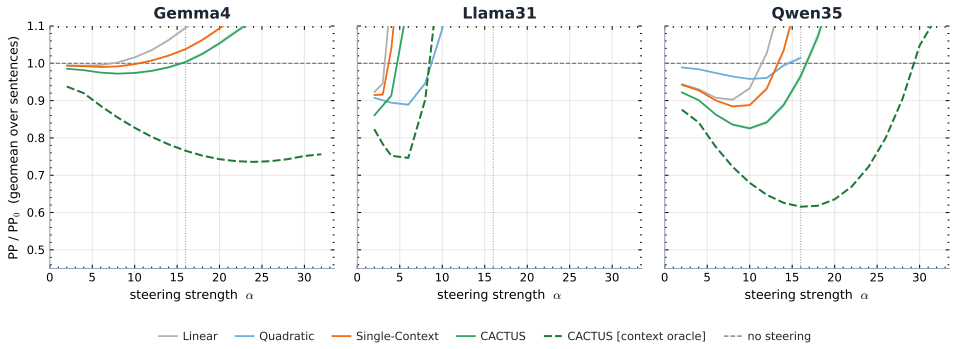

Figure 5: Mean perplexity reduction for different steering strengths α, steering language (MTOP). The lower the better.

6 Conclusion

Discussion We demonstrate that concept cannot be linear through a transformer block if it actually computes with it. This finding questions how interpretability research generally models concepts, as most works model concepts as single directions. To confirm this theoretical result, we empirically characterize this non-linearity. Concepts have high variability across contexts, and are formed of multiple different directions that form highly multidimensional structures. To effectively use those different contexts, we introduce CACTUS, a novel steering method that combines contextual concepts. CACTUS outperforms baselines on all tested LLMs, reflecting capability to efficiently insert concepts in neural activations

Limitations While CACTUS is efficient in the general case of categorical concepts, it does not consider specificities, or structures that individual concepts may have, such as sequential concepts (Wurgaft et al., 2026), or hierarchies (Park et al., 2025). Also, our arguments for concepts not being linear are tied to specific architectures.Therefore, a potential architecture with linear concepts at every point could possibly be introduced in the future. Lastly, our experiments consider supervised concepts only, while unsupervised methods such as sparse autoencoders have received massive interest from the scientific community.

Perspectives As CACTUS improves improves steering, it may be used to control the behavior of neural networks. In particular, it could be applied to inhibit sensitive biases models may have, in order to allow safer use of AI. Our theoretical and empirical results show the necessity to go beyond linear representations for concepts. It could therefore pave the way for a new category of mechanistic interpretability techniques that model concepts using contextual information, to gain better understanding and control over neural networks.

AI USE STATEMENT

Generative AI tools have been used in this work solely to aid or polish writing and scientific plotting.

ETHICS STATEMENT

This paper presents work aimed at advancing the field of mechanistic interpretability. It could lead to a better understanding and more transparent use of neural network predictions. As a method allowing to perform steering, it can be used to mitigate the biases of models and help to align it on human values. There is nevertheless no guarantee that the values on which a model is aligned by one user correspond to the desirable values of another user. For example, steering the concept of "anger" may lead to harmful content, that we did not include in provided steering examples (table 11).

REPRODUCIBILITY STATEMENT

Models and datasets used in this study are all available publicly and described in section E, most of them being available on the Huggingface Hub. Implementation details are described in section D. Code necessary to reproduce our experiments is provided as a supplementary material.

REFERENCES

Hiba Ahsan and Byron Wallace. Can saes reveal and mitigate racial biases of llms in healthcare? In *International Conference on Learning Representations*, volume 2026, pp. 4154–4172, 2026.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, 2024. ISSN 0016-0032.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024.

H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4): 377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8440–8451, 2020.

Clément Cornet, Romaric Besançon, and Hervé Le Borgne. Explaining how visual, textual and multimodal encoders share concepts. *arXiv preprint arXiv:2507.18512*, 2025.

Clement Cornet, Romaric Besançon, and Hervé Le Borgne. The deleuzian representation hypothesis. In *ICLR 2026-Fourteenth International Conference on Learning Representations*, 2026.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba E Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Róbert Csordás, Christopher Potts, Christopher D Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 248–262, 2024.

Ferdinand de Saussure, Albert Sechehaye, and Albert Riedlinger. *Cours de linguistique générale*. Payot, 1916.

G. Deleuze. *Différence et répétition*. Bibliothèque de philosophie contemporaine : Histoire de la philosophie et philosophie générale. Presses Universitaires de France, 1968. ISBN 9782130585299. URL https://books.google.gm/books?id=8lEwAAAAYAAJ.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4040–4054, 2020.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025.

Thomas Fel, Ekdeep Singh Lubana, Jacob S Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E Ba, and Talia Konkle. Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Forty-second International Conference on Machine Learning (ICLR)*, 2025.

Sheridan Feucht, Tal Haklay, Usha Bhalla, Daniel Wurgaft, Can Rager, Jack Merullo, Thomas McGrath, Raphaël Sarfati, Owen Lewis, Ekdeep Singh Lubana, Thomas Fel, and Atticus Geiger. Arithmetic in the wild: Llama uses base-10 addition to reason about cyclic concepts. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=8rId8lL5zY.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Representation Learning (ICLR)*, 2025.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *Interspeech 2021*, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 12–21, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021.

Georg Wilhelm Friedrich Hegel. *Wissenschaft der Logik: Die objective Logik*, volume 2. Johann Leonhard Schrag, 1816.

Oscar Hill, Mateo Espinosa Zarlenga, and Mateja Jamnik. Hierarchical concept-based interpretable models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1383–1391, 2015.

Faaiz Joad, Majd Hawasly, Sabri Boughorbel, Nadir Durrani, and Husrev Taha Sencar. There is more to refusal in large language models than a single direction. *arXiv preprint arXiv:2602.02132*, 2026.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5338–5348, 2020.

George Lakoff. *Women, Fire and Dangerous Thing: What Categories Reveal About the Mind*. University of Chicago Press, 1987.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025.

Matthew Levinson. Metasaes: Joint training with a decomposability penalty produces more atomic sparse autoencoder latents. *arXiv preprint arXiv:2604.03436*, 2026.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2950–2962, 2021.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11999–12009. IEEE, 2022a.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022b.

Charles Lovering, Jessica Forde, George Konidaris, Ellie Pavlick, and Michael Littman. Evaluation beyond task performance: analyzing concepts in alphazero in hex. *Advances in neural information processing systems*, 35:25992–26006, 2022.

Théo Mariotte, Martin Lebourdais, Antonio Almudévar, Marie Tahon, Alfonso Ortega, and Nicolas Dugué. Sparse autoencoders make audio foundation models more explainable. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5256–5260. IEEE, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Douglas L Medin and Marguerite M Schaffer. Context theory of classification learning. *Psychological review*, 85(3):207, 1978.

Albert Miao, Chenliang Zhou, Jiawei Zhou, and Cengiz Oztireli. Features emerge as discrete states: The first application of saes to 3d representations. *arXiv preprint arXiv:2512.11263*, 2025.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

Aayush Mishra, Daniel Khashabi, and Anqi Liu. Steered llm activations are non-surjective. *arXiv preprint arXiv:2604.09839*, 2026.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=DaNnkQJSQf.

Jayneel Parekh, Pegah KHAYATAN, Mustafa Shukor, Arnaud Dapogny, Alasdair Newson, and Matthieu Cord. Learning to steer: Input-dependent steering for multimodal LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QSLlPljXxz.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 39643–39666, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415. IEEE, 2019.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Stefano Recanatesi, Serena Bradde, Vijay Balasubramanian, Nicholas A Steinmetz, and Eric Shea-Brown. A scale-dependent measure of system dimensionality. *Patterns*, 3(8), 2022.

Eleanor Rosch. Principles of categorization. In *Cognition and Categorization*. Lawrence Erlbaum, 1978.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, et al. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Interpretable and testable vision features via sparse autoencoders. *arXiv preprint arXiv:2502.06755*, 2025.

Jiuding Sun, Sidharth Baskaran, Zhengxuan Wu, Michael Sklar, Christopher Potts, and Atticus Geiger. Hypersteer: Activation steering at scale with hypernetworks. *arXiv preprint arXiv:2506.03292*, 2025.

Wei Ren Tan, Chee Seng Chan, Hernan E Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1): 394–409, 2018.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. *arXiv preprint arXiv:2502.03714*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2025. URL https://openreview.net/forum?id=2XBPdPIcFK.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Towards universality: Studying mechanistic similarity across language model architectures. In *International Conference on Learning Representations*, volume 2025, pp. 91282–91308, 2025.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962, 2024.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=K2CckZjNy0.

Daniel Wurgaft, Can Rager, Matthew Kowal, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphaël Sarfati, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=pf3HKAGT2M.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting clip with hierarchical sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Dylan Zhou, Kunal Patil, Yifan Sun, Karthik lakshmanan, Senthooran Rajamanoharan, and Arthur Conmy. LLM neurosurgeon: Targeted knowledge removal in LLMs using sparse autoencoders. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. URL https://openreview.net/forum?id=aeQeXlG2Pw.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924. IEEE, 2021.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Appendix: Empirical Characterization

We provide additional, and detailed results for our main experiments. In particular, we provide numerical results, that are not aggregated upon models.

A.1 Diversity of concepts

Tables 2 to 4 contain results for mean pairwise cosines of contextual concepts (corresponding to section 4.1). They show per-model diversity, at both intermediate and last layer of models. Similar analysis of non-uniformity can be drawn at both stages. The most significant "model-specific" behavior concerns XlmRoberta for language attributes, that reach median cosines of at least 0.88 for both MTOP and MASSIVE, at intermermediate and last layers.

A.2 Diversity under ZCA whitening

As our linear concept extraction technique is not Fisher-optimal, we measure the diversity of contextual concepts after applying ZCA whitening to the activations, in order to account for covariance structures. Results are presented in tables 5 to 7. Those results give median cosines typically below 0.3, and still highlight a high variability of contextual concepts.

Table 2: Image modality: alignment of the contextual directions of the same concept, per model. Each entry is a quantile of the cosine $\cos(\mathrm{v}_c^{(k)}, \mathrm{v}_c^{(k')})$ between the contextual directions of a concept c in two contexts k, k' (q_{50} is the median), at intermediate and last layer.

			intermediate layer					last layer				
Dataset	Concept	Model	q_5	q_{25}	q_{50}	q_{75}	q_{95}	q_5	q_{25}	q_{50}	q_{75}	q_{95}
DomainNet	Domain	DinoV3	0.54	0.74	0.84	0.90	0.95	0.20	0.34	0.45	0.55	0.66
		OpenClipVision	0.55	0.78	0.88	0.94	0.97	0.37	0.50	0.60	0.69	0.78
		SwinV2	0.31	0.52	0.65	0.74	0.83	0.02	0.46	0.64	0.77	0.88
		ResNet50	0.37	0.65	0.78	0.87	0.93	0.29	0.45	0.57	0.68	0.80
		ConvNext	0.22	0.54	0.73	0.84	0.91	0.22	0.40	0.52	0.66	0.81
DomainNet	Label	DinoV3	-0.06	0.13	0.29	0.47	0.67	0.17	0.41	0.63	0.80	0.90
		OpenClipVision	-0.08	0.12	0.30	0.50	0.73	0.16	0.39	0.61	0.77	0.88
		SwinV2	0.13	0.34	0.53	0.69	0.83	-0.08	0.19	0.43	0.65	0.84
		ResNet50	-0.05	0.19	0.39	0.57	0.77	0.01	0.18	0.41	0.62	0.80
		ConvNext	-0.13	0.22	0.44	0.62	0.81	-0.02	0.16	0.44	0.71	0.91
WikiArt	Genre	DinoV3	-0.14	0.11	0.33	0.55	0.78	-0.01	0.20	0.40	0.60	0.80
		OpenClipVision	-0.19	0.09	0.32	0.57	0.80	0.01	0.22	0.42	0.62	0.80
		SwinV2	-0.07	0.16	0.37	0.61	0.81	-0.26	0.13	0.40	0.67	0.89
		ResNet50	-0.19	0.09	0.32	0.58	0.80	-0.07	0.10	0.31	0.55	0.78
		ConvNext	-0.37	0.08	0.42	0.66	0.86	-0.10	0.08	0.26	0.51	0.76
WikiArt	Style	DinoV3	-0.08	0.29	0.49	0.71	0.87	0.04	0.23	0.42	0.60	0.78
		OpenClipVision	-0.08	0.19	0.44	0.63	0.85	0.05	0.35	0.55	0.69	0.83
		SwinV2	-0.03	0.23	0.43	0.60	0.79	-0.25	0.12	0.34	0.52	0.73
		ResNet50	-0.09	0.22	0.43	0.65	0.86	-0.04	0.14	0.32	0.49	0.73
		ConvNext	-0.30	0.18	0.44	0.68	0.87	-0.06	0.11	0.26	0.42	0.66
MIT States	Noun	DinoV3	-0.15	0.13	0.36	0.58	0.81	0.00	0.16	0.32	0.50	0.73
		OpenClipVision	-0.16	0.13	0.36	0.57	0.79	0.02	0.23	0.39	0.56	0.75
		SwinV2	-0.00	0.21	0.39	0.56	0.77	-0.18	0.15	0.38	0.57	0.79
		ResNet50	-0.15	0.15	0.38	0.59	0.81	-0.03	0.18	0.37	0.56	0.77
		ConvNext	-0.25	0.13	0.40	0.63	0.84	-0.05	0.16	0.39	0.60	0.83
MIT States	Adj	DinoV3	-0.24	-0.04	0.11	0.27	0.55	-0.07	0.00	0.06	0.14	0.37
		OpenClipVision	-0.25	-0.04	0.12	0.30	0.57	-0.08	0.02	0.09	0.20	0.43
		SwinV2	-0.11	0.01	0.10	0.20	0.43	-0.42	-0.09	0.10	0.31	0.62
		ResNet50	-0.25	-0.04	0.11	0.29	0.56	-0.12	-0.01	0.06	0.16	0.37
		ConvNext	-0.40	-0.10	0.12	0.33	0.61	-0.14	-0.03	0.04	0.13	0.35

Table 3: Text modality: alignment of the contextual directions of the same concept, per model. Each entry is a quantile of the cosine $\cos(\mathrm{v}_c^{(k)}, \mathrm{v}_c^{(k')})$ between the contextual directions of a concept c in two contexts k, k' (q_{50} is the median), at intermediate and last layer.

			intermediate layer					last layer				
Dataset	Concept	Model	q_5	q_{25}	q_{50}	q_{75}	q_{95}	q_5	q_{25}	q_{50}	q_{75}	q_{95}
MASSIVE	Language	XlmRoberta	0.95	0.96	0.97	0.98	0.99	0.85	0.90	0.93	0.95	0.97
		MDebertaV3	-0.03	0.61	0.80	0.90	0.95	0.28	0.63	0.77	0.85	0.92
		Gemma4	0.46	0.63	0.73	0.82	0.89	0.81	0.87	0.90	0.93	0.96
		Qwen35	0.51	0.64	0.72	0.79	0.85	0.89	0.94	0.95	0.97	0.98
		Llama31	0.68	0.75	0.80	0.84	0.89	0.86	0.92	0.94	0.96	0.97
MASSIVE	Scenario	XlmRoberta	0.05	0.21	0.36	0.48	0.67	-0.01	0.32	0.55	0.68	0.79
		MDebertaV3	-0.41	0.31	0.62	0.78	0.94	0.23	0.56	0.73	0.84	0.94
		Gemma4	0.50	0.67	0.81	0.89	0.95	0.25	0.45	0.53	0.62	0.79
		Qwen35	0.47	0.71	0.91	0.94	0.97	0.39	0.56	0.66	0.73	0.82
		Llama31	0.64	0.77	0.87	0.92	0.96	0.36	0.55	0.66	0.73	0.82
MTOP	Language	XlmRoberta	0.94	0.96	0.98	0.99	0.99	0.65	0.81	0.88	0.93	0.96
		MDebertaV3	-0.49	0.42	0.76	0.90	0.96	0.50	0.69	0.82	0.90	0.95
		Gemma4	0.42	0.62	0.76	0.84	0.93	0.77	0.84	0.89	0.92	0.96
		Qwen35	0.51	0.63	0.70	0.77	0.84	0.87	0.92	0.95	0.96	0.97
		Llama31	0.58	0.69	0.74	0.80	0.86	0.79	0.87	0.92	0.94	0.97
MTOP	Domain	XlmRoberta	0.03	0.20	0.32	0.50	0.66	0.12	0.34	0.55	0.75	0.87
		MDebertaV3	-0.55	0.13	0.52	0.74	0.92	0.23	0.51	0.73	0.81	0.88
		Gemma4	0.45	0.70	0.79	0.86	0.90	0.31	0.45	0.51	0.57	0.70
		Qwen35	0.58	0.69	0.84	0.91	0.93	0.38	0.52	0.63	0.69	0.73
		Llama31	0.67	0.76	0.81	0.86	0.89	0.36	0.51	0.63	0.68	0.75
GoEmotions	Subreddit	XlmRoberta	-0.51	-0.21	0.06	0.31	0.60	-0.72	-0.32	0.07	0.45	0.78
		MDebertaV3	-0.84	-0.49	0.06	0.57	0.87	-0.58	-0.25	0.05	0.34	0.66
		Gemma4	-0.30	-0.10	0.04	0.20	0.42	-0.38	-0.11	0.09	0.29	0.55
		Qwen35	-0.14	-0.04	0.04	0.12	0.26	-0.32	-0.09	0.10	0.28	0.51
		Llama31	-0.13	-0.03	0.05	0.14	0.28	-0.32	-0.08	0.10	0.28	0.52
GoEmotions	Emotion	XlmRoberta	-0.48	-0.18	0.08	0.33	0.60	-0.70	-0.31	0.07	0.44	0.77
		MDebertaV3	-0.83	-0.44	0.12	0.61	0.88	-0.56	-0.21	0.08	0.36	0.66
		Gemma4	-0.27	-0.08	0.07	0.22	0.44	-0.38	-0.11	0.09	0.29	0.55
		Qwen35	-0.11	-0.01	0.07	0.18	0.37	-0.31	-0.08	0.09	0.27	0.50
		Llama31	-0.11	-0.01	0.08	0.19	0.39	-0.31	-0.09	0.08	0.26	0.49

Table 4: Audio modality: alignment of the contextual directions of the same concept, per model. Each entry is a quantile of the cosine $\cos(\mathrm{v}_c^{(k)}, \mathrm{v}_c^{(k')})$ between the contextual directions of a concept c in two contexts k, k' (q_{50} is the median), at intermediate and last layer.

			intermediate layer					last layer				
Dataset	Concept	Model	q_5	q_{25}	q_{50}	q_{75}	q_{95}	q_5	q_{25}	q_{50}	q_{75}	q_{95}
CREMA-D	Label	Ast	0.62	0.79	0.85	0.93	0.95	0.48	0.68	0.77	0.84	0.90
		WavLm	0.65	0.79	0.87	0.92	0.95	0.41	0.74	0.82	0.90	0.95
		Clap	0.57	0.76	0.84	0.94	0.96	0.45	0.69	0.80	0.90	0.94
		Hubert	0.56	0.77	0.85	0.90	0.94	0.54	0.83	0.89	0.93	0.96
		Whisper	0.30	0.96	0.99	1.00	1.00	0.45	0.93	0.98	0.99	1.00
CREMA-D	Sentence	Ast	0.60	0.74	0.82	0.87	0.91	0.49	0.68	0.76	0.82	0.88
		WavLm	0.70	0.85	0.89	0.92	0.95	0.67	0.86	0.91	0.94	0.96
		Clap	0.59	0.74	0.82	0.86	0.91	0.53	0.74	0.82	0.86	0.92
		Hubert	0.80	0.88	0.92	0.94	0.96	0.83	0.92	0.96	0.98	0.99
		Whisper	-0.11	0.72	0.90	0.96	0.99	0.06	0.72	0.89	0.95	0.98
ESD	Emotion	Ast	0.03	0.35	0.51	0.63	0.76	0.01	0.33	0.51	0.64	0.79
		WavLm	0.20	0.41	0.51	0.59	0.68	0.04	0.34	0.48	0.57	0.70
		Clap	-0.02	0.34	0.52	0.64	0.76	0.10	0.34	0.49	0.61	0.73
		Hubert	0.07	0.36	0.49	0.62	0.73	-0.27	0.18	0.42	0.65	0.85
		Whisper	-0.83	0.07	0.63	0.90	0.98	-0.59	0.11	0.56	0.81	0.96
ESD	Speaker	Ast	0.15	0.40	0.63	0.74	0.84	0.06	0.48	0.66	0.77	0.85
		WavLm	0.33	0.53	0.69	0.75	0.83	0.19	0.50	0.62	0.74	0.86
		Clap	0.19	0.54	0.76	0.83	0.87	0.24	0.59	0.78	0.84	0.89
		Hubert	0.18	0.44	0.64	0.74	0.82	-0.13	0.35	0.58	0.75	0.90
		Whisper	-0.64	0.37	0.93	0.99	0.99	-0.53	0.26	0.91	0.98	0.99
AudioMNIST	Digit	Ast	0.20	0.48	0.61	0.72	0.85	0.08	0.34	0.50	0.64	0.78
		WavLm	0.61	0.72	0.79	0.85	0.91	0.69	0.80	0.86	0.90	0.94
		Clap	0.26	0.54	0.65	0.73	0.83	0.15	0.39	0.56	0.68	0.81
		Hubert	0.57	0.71	0.78	0.85	0.92	0.66	0.83	0.90	0.93	0.97
		Whisper	-0.62	0.20	0.79	0.96	0.99	-0.54	0.23	0.74	0.92	0.97
AudioMNIST	Accent	Ast	0.26	0.48	0.58	0.69	0.79	0.16	0.45	0.58	0.70	0.84
		WavLm	0.14	0.30	0.40	0.48	0.61	0.11	0.25	0.35	0.45	0.59
		Clap	0.33	0.48	0.59	0.67	0.77	0.28	0.48	0.60	0.70	0.82
		Hubert	0.05	0.24	0.35	0.46	0.61	-0.18	0.16	0.40	0.58	0.76
		Whisper	-0.85	-0.14	0.63	0.91	0.97	-0.75	-0.02	0.53	0.86	0.96

Table 5: Image modality: alignment of the contextual directions of the same concept, per model (ZCA-whitened activations). Each entry is a quantile of the cosine $\cos(\mathrm{v}_c^{(k)}, \mathrm{v}_c^{(k')})$ between the contextual directions of a concept c in two contexts k, k' (q_{50} is the median), at the half and last layer.

			half layer					last layer				
Dataset	Concept	Model	q_5	q_{25}	q_{50}	q_{75}	q_{95}	q_5	q_{25}	q_{50}	q_{75}	q_{95}
DomainNet	Domain	DinoV3	-0.01	0.06	0.13	0.21	0.35	-0.02	0.02	0.05	0.10	0.20
		OpenClipVision	-0.01	0.06	0.13	0.21	0.34	-0.01	0.03	0.07	0.12	0.22
		SwinV2	-0.02	0.03	0.08	0.14	0.23	-0.02	0.01	0.04	0.09	0.18
		ResNet50	-0.01	0.03	0.09	0.16	0.27	-0.01	0.01	0.03	0.07	0.15
		ConvNext	-0.01	0.05	0.12	0.20	0.33	-0.01	0.01	0.05	0.10	0.19
DomainNet	Label	DinoV3	0.00	0.06	0.14	0.26	0.46	0.06	0.24	0.49	0.68	0.84
		OpenClipVision	0.00	0.07	0.16	0.28	0.52	0.07	0.24	0.44	0.63	0.81
		SwinV2	0.04	0.18	0.36	0.57	0.78	0.02	0.15	0.37	0.60	0.79
		ResNet50	-0.00	0.05	0.14	0.28	0.52	0.00	0.05	0.19	0.41	0.68
		ConvNext	0.02	0.13	0.26	0.43	0.66	0.01	0.09	0.31	0.60	0.84
WikiArt	Genre	DinoV3	-0.03	0.02	0.06	0.14	0.34	-0.02	0.03	0.08	0.17	0.41
		OpenClipVision	-0.03	0.02	0.06	0.15	0.36	-0.02	0.02	0.07	0.17	0.41
		SwinV2	-0.02	0.02	0.07	0.16	0.39	-0.02	0.01	0.06	0.14	0.35
		ResNet50	-0.02	0.01	0.05	0.11	0.30	-0.02	0.00	0.03	0.09	0.25
		ConvNext	-0.03	0.02	0.08	0.18	0.40	-0.02	0.01	0.05	0.13	0.33
WikiArt	Style	DinoV3	-0.01	0.04	0.11	0.26	0.48	-0.01	0.04	0.11	0.22	0.44
		OpenClipVision	-0.01	0.04	0.11	0.24	0.46	-0.00	0.05	0.12	0.24	0.46
		SwinV2	-0.01	0.04	0.10	0.22	0.42	-0.01	0.03	0.09	0.19	0.40
		ResNet50	-0.01	0.03	0.09	0.22	0.43	-0.01	0.02	0.05	0.13	0.32
		ConvNext	-0.01	0.05	0.13	0.27	0.51	-0.02	0.03	0.08	0.19	0.39
MIT States	Noun	DinoV3	-0.02	0.04	0.10	0.19	0.39	-0.00	0.08	0.20	0.35	0.59
		OpenClipVision	-0.02	0.04	0.10	0.20	0.39	-0.00	0.09	0.20	0.36	0.59
		SwinV2	-0.01	0.07	0.17	0.30	0.53	-0.01	0.08	0.19	0.33	0.58
		ResNet50	-0.02	0.03	0.09	0.19	0.38	-0.01	0.03	0.10	0.21	0.44
		ConvNext	-0.01	0.07	0.16	0.30	0.53	-0.01	0.06	0.16	0.32	0.61
MIT States	Adj	DinoV3	-0.05	-0.01	0.02	0.05	0.15	-0.05	-0.01	0.02	0.06	0.21
		OpenClipVision	-0.05	-0.01	0.02	0.06	0.15	-0.04	-0.00	0.03	0.07	0.21
		SwinV2	-0.04	-0.01	0.02	0.06	0.17	-0.04	-0.01	0.02	0.05	0.17
		ResNet50	-0.04	-0.01	0.01	0.05	0.14	-0.03	-0.01	0.01	0.03	0.11
		ConvNext	-0.06	-0.01	0.03	0.07	0.20	-0.04	-0.01	0.01	0.04	0.16

Table 6: Text modality: alignment of the contextual directions of the same concept, per model (ZCA-whitened activations). Each entry is a quantile of the cosine $\cos(\mathrm{v}_c^{(k)}, \mathrm{v}_c^{(k')})$ between the contextual directions of a concept c in two contexts k, k' (q_{50} is the median), at the half and last layer.

			half layer					last layer				
Dataset	Concept	Model	q_5	q_{25}	q_{50}	q_{75}	q_{95}	q_5	q_{25}	q_{50}	q_{75}	q_{95}
MASSIVE	Language	XlmRoberta	0.06	0.12	0.17	0.25	0.34	0.08	0.14	0.19	0.27	0.37
		MDebertaV3	0.10	0.17	0.23	0.30	0.42	0.16	0.25	0.33	0.41	0.52
		Gemma4	0.10	0.15	0.20	0.27	0.36	-0.01	0.02	0.06	0.10	0.21
		Qwen35	0.06	0.11	0.16	0.23	0.32	-0.01	0.01	0.04	0.08	0.17
		Llama31	0.04	0.09	0.14	0.21	0.31	-0.01	0.02	0.04	0.08	0.17
MASSIVE	Scenario	XlmRoberta	0.04	0.10	0.17	0.25	0.37	0.07	0.15	0.23	0.32	0.44
		MDebertaV3	0.12	0.20	0.29	0.39	0.52	0.17	0.25	0.39	0.51	0.64
		Gemma4	0.17	0.27	0.39	0.49	0.61	0.03	0.05	0.07	0.10	0.16
		Qwen35	0.07	0.16	0.33	0.43	0.57	0.03	0.04	0.06	0.09	0.13
		Llama31	0.10	0.16	0.24	0.35	0.48	0.02	0.04	0.06	0.08	0.15
MTOP	Language	XlmRoberta	0.03	0.06	0.09	0.14	0.21	0.06	0.09	0.13	0.18	0.25
		MDebertaV3	0.06	0.09	0.13	0.18	0.28	0.12	0.19	0.23	0.30	0.41
		Gemma4	0.08	0.13	0.16	0.21	0.31	-0.00	0.02	0.04	0.07	0.17
		Qwen35	0.04	0.08	0.11	0.15	0.23	-0.00	0.01	0.03	0.05	0.13
		Llama31	0.03	0.06	0.09	0.13	0.20	0.00	0.01	0.03	0.05	0.13
MTOP	Domain	XlmRoberta	0.03	0.06	0.09	0.16	0.20	0.05	0.09	0.14	0.19	0.27
		MDebertaV3	0.05	0.11	0.16	0.23	0.32	0.13	0.19	0.26	0.35	0.47
		Gemma4	0.14	0.21	0.27	0.35	0.47	0.02	0.04	0.05	0.07	0.09
		Qwen35	0.05	0.10	0.19	0.29	0.39	0.02	0.03	0.04	0.05	0.08
		Llama31	0.07	0.10	0.14	0.19	0.29	0.01	0.03	0.04	0.05	0.08
GoEmotions	Subreddit	XlmRoberta	-0.05	-0.02	0.01	0.03	0.07	-0.05	-0.01	0.01	0.03	0.07
		MDebertaV3	-0.05	-0.02	0.01	0.03	0.07	-0.06	-0.02	0.00	0.03	0.07
		Gemma4	-0.03	-0.01	0.00	0.02	0.04	-0.03	-0.01	0.01	0.02	0.07
		Qwen35	-0.02	-0.01	0.00	0.01	0.03	-0.02	-0.01	0.00	0.02	0.05
		Llama31	-0.02	-0.01	0.00	0.02	0.04	-0.02	-0.01	0.00	0.02	0.05
GoEmotions	Emotion	XlmRoberta	-0.04	-0.01	0.01	0.04	0.08	-0.04	-0.01	0.01	0.04	0.08
		MDebertaV3	-0.05	-0.01	0.02	0.05	0.10	-0.05	-0.01	0.01	0.04	0.09
		Gemma4	-0.03	-0.01	0.00	0.02	0.04	-0.03	-0.01	0.01	0.02	0.05
		Qwen35	-0.02	-0.01	0.00	0.02	0.03	-0.02	-0.01	0.00	0.02	0.04
		Llama31	-0.02	-0.01	0.00	0.02	0.04	-0.02	-0.01	0.00	0.02	0.04

Table 7: Audio modality: alignment of the contextual directions of the same concept, per model (ZCA-whitened activations). Each entry is a quantile of the cosine $\cos(\mathrm{v}_c^{(k)}, \mathrm{v}_c^{(k')})$ between the contextual directions of a concept c in two contexts k, k' (q_{50} is the median), at the half and last layer.

			half layer					last layer				
Dataset	Concept	Model	q_5	q_{25}	q_{50}	q_{75}	q_{95}	q_5	q_{25}	q_{50}	q_{75}	q_{95}
CREMA-D	Label	Ast	0.03	0.07	0.11	0.14	0.20	0.03	0.08	0.11	0.15	0.20
		WavLm	0.02	0.05	0.08	0.10	0.14	0.01	0.04	0.06	0.08	0.11
		Clap	0.06	0.13	0.18	0.22	0.29	0.02	0.07	0.10	0.13	0.20
		Hubert	0.01	0.04	0.06	0.08	0.11	0.01	0.03	0.04	0.06	0.09
		Whisper	0.02	0.05	0.07	0.10	0.14	0.02	0.04	0.06	0.09	0.13
CREMA-D	Sentence	Ast	0.15	0.29	0.34	0.40	0.46	0.15	0.27	0.32	0.37	0.44
		WavLm	0.12	0.20	0.27	0.36	0.44	0.09	0.16	0.25	0.35	0.41
		Clap	0.27	0.42	0.48	0.56	0.61	0.14	0.26	0.32	0.39	0.45
		Hubert	0.10	0.17	0.23	0.33	0.40	0.07	0.15	0.22	0.31	0.38
		Whisper	0.11	0.20	0.26	0.34	0.43	0.10	0.17	0.25	0.33	0.42
ESD	Emotion	Ast	-0.02	0.01	0.05	0.09	0.14	-0.01	0.02	0.06	0.10	0.17
		WavLm	-0.00	0.02	0.04	0.06	0.10	-0.00	0.04	0.07	0.11	0.17
		Clap	-0.03	0.02	0.07	0.12	0.19	-0.01	0.00	0.03	0.07	0.14
		Hubert	-0.01	0.02	0.04	0.06	0.09	0.00	0.04	0.07	0.10	0.15
		Whisper	-0.01	0.02	0.04	0.07	0.12	-0.01	0.02	0.04	0.06	0.11
ESD	Speaker	Ast	0.03	0.08	0.16	0.23	0.34	0.03	0.09	0.18	0.23	0.36
		WavLm	0.03	0.07	0.12	0.17	0.30	0.03	0.11	0.20	0.26	0.41
		Clap	0.06	0.13	0.22	0.31	0.45	0.02	0.06	0.14	0.21	0.36
		Hubert	0.02	0.06	0.12	0.17	0.27	0.04	0.11	0.19	0.25	0.39
		Whisper	0.02	0.07	0.12	0.19	0.28	0.02	0.06	0.12	0.18	0.28
AudioMNIST	Digit	Ast	-0.02	0.02	0.07	0.13	0.23	-0.01	0.03	0.07	0.14	0.25
		WavLm	-0.03	0.01	0.04	0.09	0.16	-0.04	0.00	0.03	0.09	0.17
		Clap	-0.02	0.04	0.10	0.17	0.27	-0.02	0.01	0.05	0.10	0.20
		Hubert	-0.03	0.01	0.04	0.08	0.15	-0.03	0.01	0.04	0.09	0.17
		Whisper	-0.03	0.00	0.03	0.07	0.13	-0.03	0.00	0.03	0.07	0.14
AudioMNIST	Accent	Ast	0.01	0.07	0.13	0.18	0.30	0.01	0.07	0.12	0.18	0.31
		WavLm	0.01	0.03	0.05	0.08	0.15	0.00	0.02	0.03	0.06	0.11
		Clap	0.03	0.10	0.16	0.22	0.33	0.00	0.05	0.09	0.15	0.25
		Hubert	0.01	0.03	0.05	0.08	0.15	0.00	0.02	0.04	0.06	0.11
		Whisper	0.00	0.02	0.04	0.07	0.13	0.00	0.02	0.04	0.07	0.13

A.3 SIZE OF CONCEPTS

Tables 2 to 4 contain detailed results for the effective dimension per concept from section 4.2. Those results are detailed per model.

Table 8: Image models, per-model effective dimension d_{eff} (mean $\pm$ std over the concept values of each dataset $\times$ direction), at intermediate and last layers.

Dataset	Concept	Contexts	Model	d_{eff} (intermediate)	d_{eff} (last)
DomainNet	domain	345	DinoV3	1.54 ± 0.28	5.46 ± 2.58
			OpenClipVision	1.45 ± 0.34	2.89 ± 0.77
			SwinV2	2.73 ± 1.06	2.79 ± 1.01
			ResNet50	1.86 ± 0.52	3.25 ± 1.11
			ConvNext	2.35 ± 1.04	3.87 ± 1.59
DomainNet	label	6	DinoV3	3.62 ± 0.59	2.06 ± 0.45
			OpenClipVision	3.45 ± 0.63	2.15 ± 0.47
			SwinV2	2.43 ± 0.44	2.74 ± 0.52
			ResNet50	3.08 ± 0.64	2.92 ± 0.56
			ConvNext	2.81 ± 0.55	2.64 ± 0.69
WikiArt	genre	27	DinoV3	5.10 ± 1.67	4.68 ± 2.28
			OpenClipVision	5.10 ± 2.07	4.51 ± 2.18
			SwinV2	5.01 ± 2.60	4.02 ± 1.74
			ResNet50	5.27 ± 2.60	5.73 ± 2.83
			ConvNext	3.80 ± 1.24	6.08 ± 2.80
WikiArt	style	11	DinoV3	2.73 ± 0.93	3.47 ± 1.13
			OpenClipVision	3.25 ± 1.20	2.72 ± 0.78
			SwinV2	3.36 ± 1.14	3.76 ± 0.96
			ResNet50	3.18 ± 1.23	4.25 ± 1.56
			ConvNext	2.91 ± 0.87	4.79 ± 1.57
MIT States	noun	113	DinoV3	3.60 ± 1.17	4.02 ± 1.39
			OpenClipVision	3.61 ± 1.14	3.55 ± 1.15
			SwinV2	3.57 ± 1.18	3.48 ± 1.08
			ResNet50	3.51 ± 1.15	3.71 ± 1.30
			ConvNext	3.28 ± 1.09	3.58 ± 1.43
MIT States	adj	245	DinoV3	7.39 ± 3.46	11.66 ± 7.43
			OpenClipVision	6.97 ± 3.18	10.05 ± 5.84
			SwinV2	9.82 ± 5.55	6.04 ± 2.50
			ResNet50	7.18 ± 3.34	10.77 ± 6.19
			ConvNext	5.84 ± 2.41	11.27 ± 6.83

Table 9: Text models, per-model effective dimension d_{eff} (mean $\pm$ std over the concept values of each dataset $\times$ direction), at intermediate and last layers.

Dataset	Concept	Contexts	Model	d_{eff} (intermediate)	d_{eff} (last)
MASSIVE	language	18	XlmRoberta	1.05 ± 0.02	1.16 ± 0.04
			MDebertaV3	1.79 ± 0.53	1.93 ± 0.77
			Gemma4	1.91 ± 0.50	1.23 ± 0.07
			Qwen35	1.90 ± 0.31	1.11 ± 0.03
			Llama31	1.54 ± 0.08	1.14 ± 0.06
MASSIVE	scenario	7	XlmRoberta	3.63 ± 0.51	2.56 ± 0.36
			MDebertaV3	2.19 ± 0.59	1.83 ± 0.36
			Gemma4	1.51 ± 0.13	2.61 ± 0.54
			Qwen35	1.34 ± 0.08	2.01 ± 0.21
			Llama31	1.33 ± 0.07	2.02 ± 0.23
MTOP	language	11	XlmRoberta	1.05 ± 0.03	1.32 ± 0.12
			MDebertaV3	1.90 ± 0.63	1.55 ± 0.33
			Gemma4	1.80 ± 0.48	1.27 ± 0.11
			Qwen35	1.92 ± 0.34	1.13 ± 0.05
			Llama31	1.71 ± 0.16	1.21 ± 0.12
MTOP	domain	5	XlmRoberta	3.24 ± 0.69	2.30 ± 0.62
			MDebertaV3	2.17 ± 0.50	1.76 ± 0.34
			Gemma4	1.50 ± 0.12	2.45 ± 0.37
			Qwen35	1.39 ± 0.07	2.05 ± 0.22
			Llama31	1.40 ± 0.08	2.02 ± 0.17
GoEmotions	subreddit	28	XlmRoberta	6.36 ± 1.15	3.97 ± 0.90
			MDebertaV3	2.87 ± 0.68	5.38 ± 0.88
			Gemma4	10.54 ± 1.84	7.75 ± 1.29
			Qwen35	15.35 ± 2.18	8.57 ± 1.43
			Llama31	14.86 ± 2.17	8.54 ± 1.44
GoEmotions	emotion	483	XlmRoberta	8.55 ± 0.77	4.50 ± 0.43
			MDebertaV3	2.97 ± 0.26	6.54 ± 0.91
			Gemma4	19.94 ± 4.56	11.63 ± 1.63
			Qwen35	45.50 ± 16.33	14.53 ± 2.76
			Llama31	43.98 ± 17.54	14.60 ± 2.70

Table 10: Audio models, per-model effective dimension d_{eff} (mean $\pm$ std over the concept values of each dataset $\times$ direction), at intermediate and last layers.

Dataset	Concept	Contexts	Model	d_{eff} (intermediate)	d_{eff} (last)
CREMA-D	label	12	Ast	1.40 ± 0.25	1.73 ± 0.40
			WavLm	1.36 ± 0.16	1.55 ± 0.33
			Clap	1.43 ± 0.30	1.63 ± 0.42
			Hubert	1.45 ± 0.25	1.33 ± 0.18
			Whisper	1.26 ± 0.56	1.25 ± 0.49
CREMA-D	sentence	6	Ast	1.45 ± 0.20	1.61 ± 0.25
			WavLm	1.25 ± 0.11	1.23 ± 0.11
			Clap	1.45 ± 0.20	1.48 ± 0.23
			Hubert	1.18 ± 0.06	1.11 ± 0.06
			Whisper	1.44 ± 0.34	1.46 ± 0.39
ESD	emotion	10	Ast	3.13 ± 0.90	3.16 ± 0.99
			WavLm	3.05 ± 0.55	3.24 ± 0.39
			Clap	3.15 ± 0.99	3.15 ± 0.78
			Hubert	3.07 ± 0.63	3.13 ± 0.90
			Whisper	1.88 ± 0.50	2.19 ± 0.49
ESD	speaker	5	Ast	2.20 ± 0.75	2.04 ± 0.68
			WavLm	1.92 ± 0.43	2.05 ± 0.54
			Clap	1.86 ± 0.73	1.77 ± 0.60
			Hubert	2.09 ± 0.54	2.23 ± 0.66
			Whisper	1.50 ± 0.42	1.59 ± 0.50
AudioMNIST	digit	17	Ast	2.72 ± 1.23	3.49 ± 1.28
			WavLm	1.59 ± 0.15	1.38 ± 0.07
			Clap	2.48 ± 0.93	3.09 ± 1.02
			Hubert	1.62 ± 0.21	1.29 ± 0.09
			Whisper	1.83 ± 0.56	2.09 ± 0.76
AudioMNIST	accent	10	Ast	2.52 ± 0.56	2.55 ± 0.63
			WavLm	4.01 ± 0.60	4.57 ± 1.07
			Clap	2.54 ± 0.54	2.44 ± 0.58
			Hubert	4.40 ± 0.71	3.78 ± 1.11
			Whisper	1.80 ± 0.38	2.12 ± 0.49

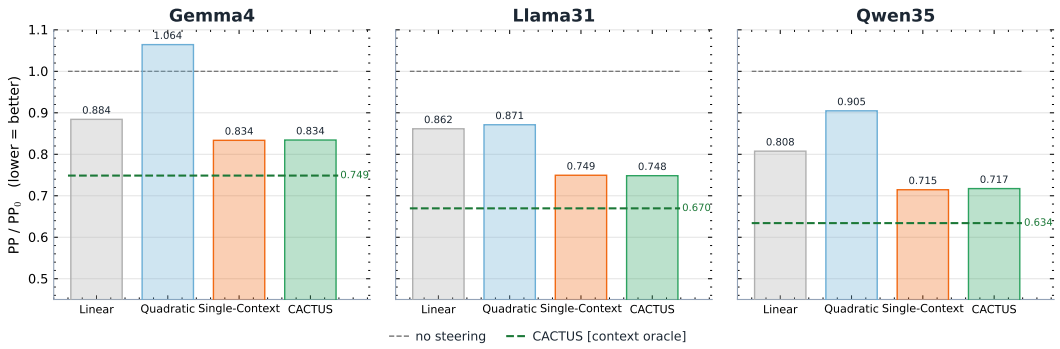

Figure 6: Perplexity reductions of different steering methods for GoEmotions emotion (intermediate layers)

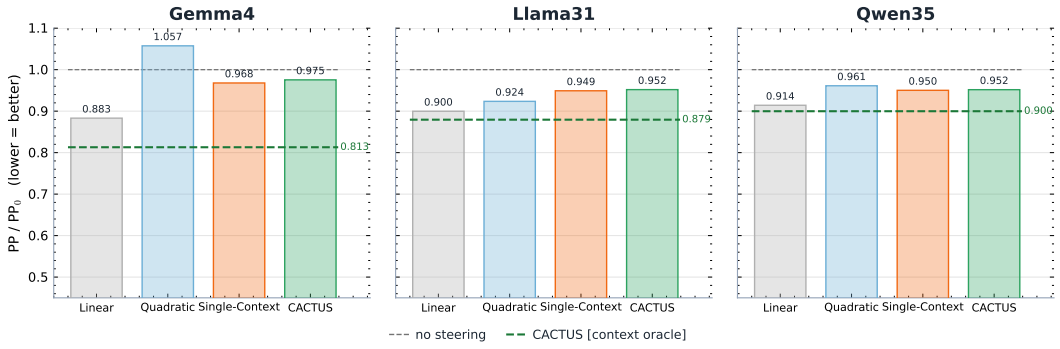

Figure 7: Perplexity reductions of different steering methods for GoEmotions emotion (last layers)

B Appendix:Steering results

B.1 Optimal Steering

As fig. 4 comprises results aggregated per-model, we provide model-wise steering results for optimal steering. Those results are provided at both intermediate and last layers of models. Every relevant steering method obtains better steering results at intermediate layers. Results are presented in figs. 6 to 11.

B.2 Steering magnitude

As results in fig. 5 display the perplexity gains only for MTOP, we provide the corresponding figures for all tested datasets, at both intermediate and last layers in figs. 12 to 16. The ordering between methods performance is almost constant across α. The only exception is quadratic steering on Qwen3.5 MASSIVE, but it is unstable, and its lowest reachable perplexity stays largely worse than other methods.

C Appendix: Qualitative examples of steering, open-ended generation

Table 11 provides qualitative examples of steering with CACTUS, on Gemma4-E4B, at intermediate layer (21st block). Each row contains 3 open-ended generation outputs: without steering, steering "joy", and steering "anger", both being extracted from GoEmotions. All 3 examples per row use the same random seeds: steering is the only thing that differentiates them.

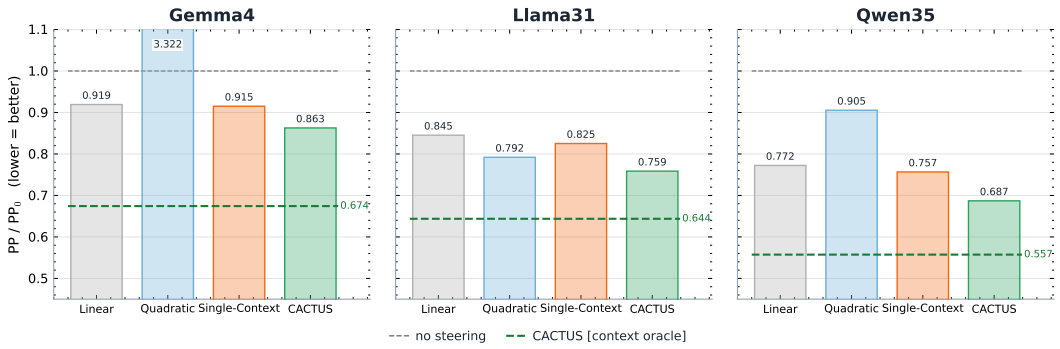

Figure 8: Perplexity reductions of different steering methods for MTOP language (intermediate layers)

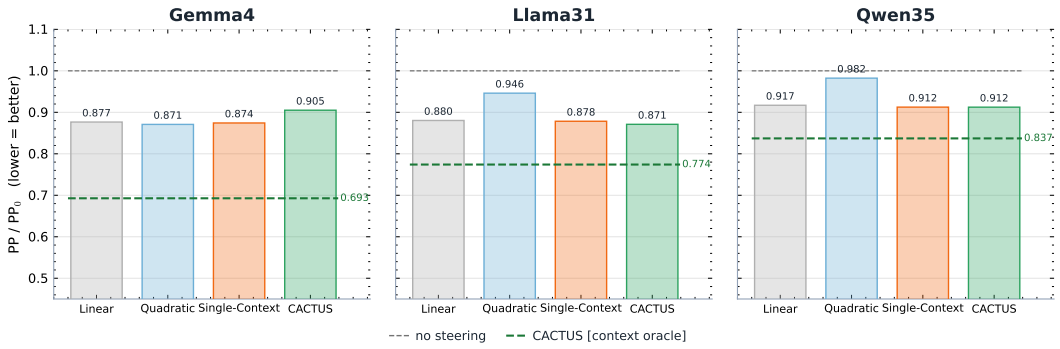

Figure 9: Perplexity reductions of different steering methods for MTOP language (last layers)

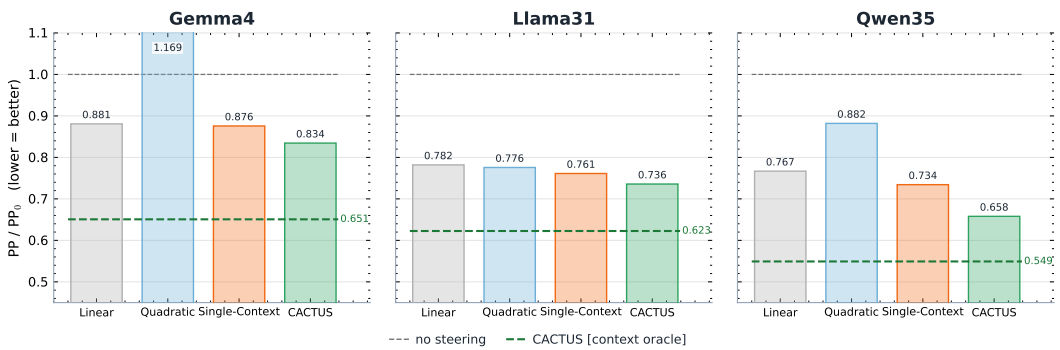

Figure 10: Perplexity reductions of different steering methods for MASSIVE language (intermediate layers)

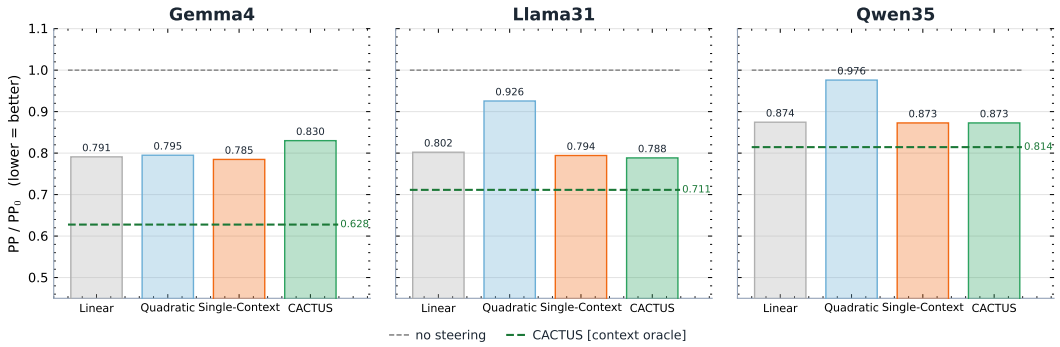

Figure 11: Perplexity reductions of different steering methods for MASSIVE language (last layers)

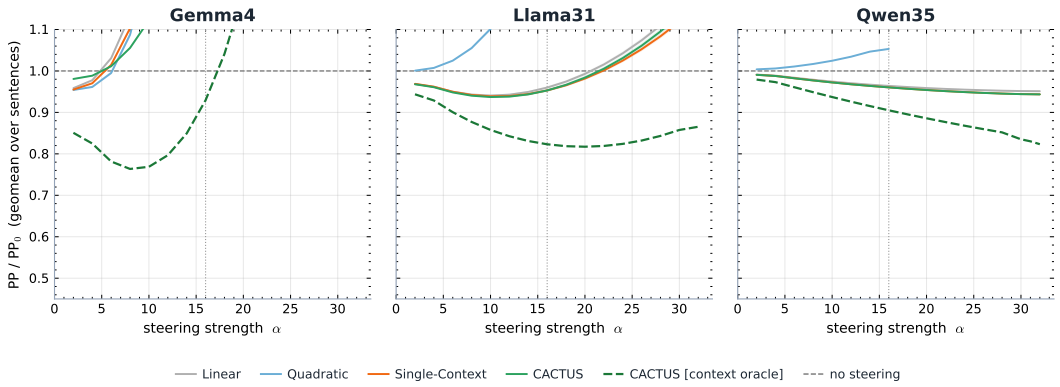

Figure 12: Perplexity reductions vs. steering magnitude for MTOP language (last layers)

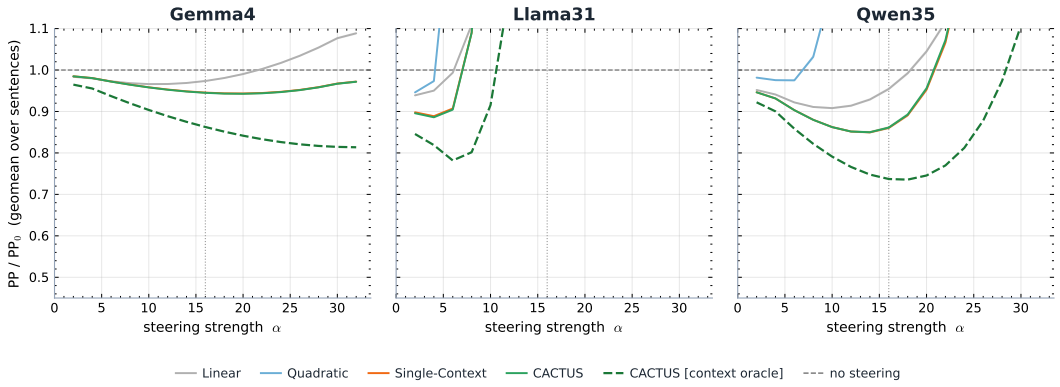

Figure 13: Perplexity reductions vs. steering magnitude for GoEmotions emotion (intermediate layers)

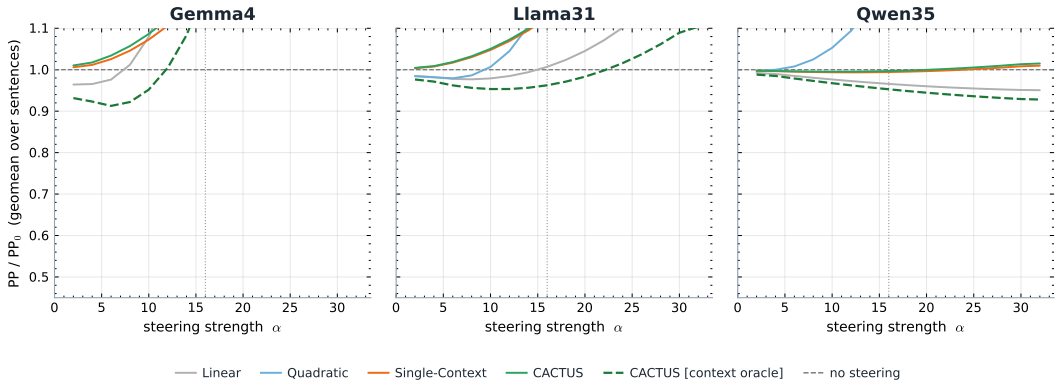

Figure 14: Perplexity reductions vs. steering magnitude for GoEmotions emotion (last layers)

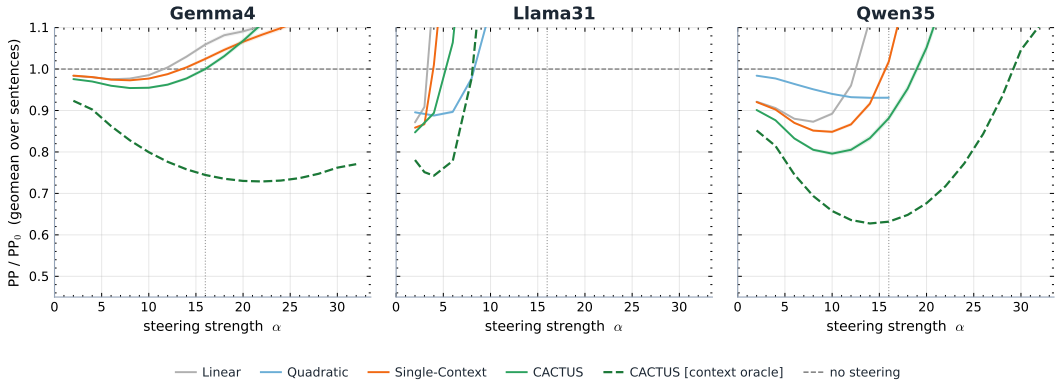

Figure 15: Perplexity reductions vs. steering magnitude for MASSIVE language (intermediate layers)

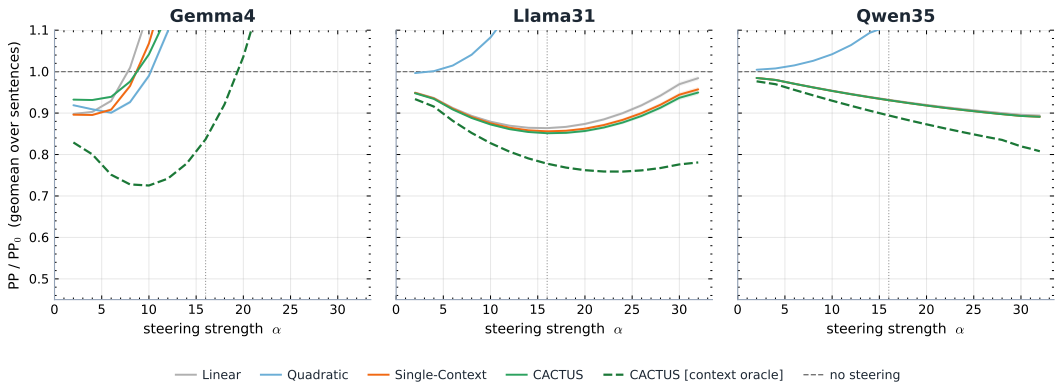

Figure 16: Perplexity reductions vs. steering magnitude for MASSIVE language (last layers)

Table 11: Steering results for "Joy" and "Anger" in Gemma4-E4B at layer 21. Rows have samples from the same random seed

No steering	**Joy** ($\alpha = 16$)	**Anger** ($\alpha = 16$)
2013 New York Consolidated Laws ISC - Insurance Article 73 - (7301 - 7323) LIFE HEALTH AND ACCIDENT INSURANCE POLICIES AND CONTRACTS 7301 - Short	2013 New York City Marathon Sunday, November 3, 2013 Thank you for joining me this year at the New York City Marathon! It was a great event, and the weather was perfect!	2013 New York Consolidated Laws PUN - Penal Part 3 - OFFENSES AGAINST THE PUBLIC Title G - OFFENSES AGAINST PUBLIC HEALTH Article 240 - (240.00 - 2}
There are several reasons why you should have a business plan to help you succeed. Having a business plan can help ensure you get the money you need to start or grow your business, give you a plan for taking your business into the future and help you	There are several reasons why you should have a strong online presence as a small business owner. With the internet growing rapidly, many people are turning to the online world to find products and services. Whether you're looking for the latest gadget or a new	There are several reasons why you should hire a plumber. You may need to make some repairs or have a plumbing emergency. Hiring a plumber can make these problems much worse. The job requires the removal of water and sewage. They're also extremely skilled
New! 18" Dollhouse Miniature Kitchen Cabinets with Working Drawers Kitchen Cabinet with Working Drawers and 2 Doors This wonderful 18" dollhouse miniature kitchen cabinet set is a great way to add a working	New! 1960 Ford Thunderbird Coupe This beautiful 1960 Ford Thunderbird has undergone a recent restoration. The original owner was a very well known and successful business man. He was	New! 18" Dollhouse Mini Kitchen Play Set with Working Range, Microwave, Sink, Refrigerator. Playset is complete with Fridge/Freezer, Microwave, Dishwasher, Oven, Cooker Stove & Sink. Comes complete with
Description * 100% Brand new. * Material:Plastic * Plastic frame size: 13.5*13.5cm * Total length(	Description * 100% Brand new. * Material:Plastic * Color:pink,blue,red * 2 kinds of shapes:heart/flower *	Description: Vigilante (1968) *The only thing you know that makes sense is fear.* * Title: Vigilante (1968) * Rating: R
The first 24 hours of COVID-19 in the US The first official case of COVID-19 in the U.S. was reported in Washington state on Jan. 21, and the Centers for	The End of 2019 ! Christmas Celebrations in the Mountains of North Carolina! The year is ending and we are enjoying one of our favorite times together with the family. The children have	The End of 2019 Is Almost Here Christmas is just around the corner, and the year is almost up. Everyone is going around yelling for 2019 to stop. Why? Because they want
Welcome to My Blog Welcome to my blog! I'm excited to share some insights and updates on various topics related to technology, digital marketing, and personal growth. Here, you'll find a mix of informative posts, creative ideas	Welcome to My Blog Hey there! I'm thrilled to welcome you to my blog! I can't wait to share with you all the latest updates and exciting things that have been happening. This blog is the perfect place to	Welcome to My Blog Welcome to my blog! I'm Michael. In this blog I will write about every aspect of technology that I feel is worthy to be written about. That's all I've got to say.
180 gram audiophile vinyl / Gatefold sleeve / INCLUDES INSERT / INCLUDES FREE DOWNLOAD CODE / LIMITED, NUMBERED EDITION OF 1000 COPIES / FORMER THE SMITHS' ROCKET SUMMER' MEMBER STE	180 gram audiograde. Gatefold sleeve. On November 19, 1981, the world premiere of this classic album was broadcast live from the Grand Ballroom of New York City's Roseland Theater. After	180 gram audiophile vinyl / FIRST TIME ON VINYL! / A RATHER QUIET PIECE OF WORK / PRODUCED BY THE ROOTS / THE ROOTS ARE BACK AGAIN! / THE THIRD ALBUM IN THE ROOTS CRIP

Steering results for "Joy" and "Anger" in Gemma4-E4B at layer 21. Rows have samples from the same random seed (Continued)

No steering	**Joy** ($\alpha = 16$)	**Anger** ($\alpha = 16$)
`[instruction]fixed all the issues that showed up when updating project [filepath]routes/html Routes.js [code]var db = require("../models"); module.exports = function(app) { // Load index page}`	`[instruction]Add tests for random [filepath]tests/test _random.py [code]"""test_rand om.py - tests."""` `import unittest from random_walker import random class RandomTest(unittest.Te stCase):`	`[instruction]fixed all the error that occur while testing my code [filepath]routes/api. php [code]<?php use Illuminate\Http\Requ est; /* |-`
Q: what is the first thing we do when assessing pain A: use pain assessment tool Q: what is an example of a pain assessment tool A: Wong-Baker Faces Pain Rating Scale Q: what is the second thing we	Q: what is the first thing we do when assessing our patient? A: ABCDEF Q: what is A-B-C-D-E-F A: a-airway b-breathing c-	Q: what is the first thing we do when assessing people with dysrhythmias? A: assess the heart rate Q: what is the normal range of the heart rate? A: 60-100 beats per
A 12-week programme of research Location: At a venue near you (see the map below) Timing: The training and programme will start in September 2013.	A 12-week programme of music Thursday 24th February 2021 4:00pm- 8:00pm Online via Zoom Join	A 12-week programme of work *The last 12 weeks have been busy. Very busy. 12 weeks ago I found myself in the middle of an argument with a 9*
A/N: The last 2 chapters will be in the perspective of Taehyung, I want to put them last to make a difference between the 2 ———————— ———————— 3rd person P.O.V	A/N: Hi, guys! Just another chapter in our quest for a better, brighter future. You already know the drill, and we won't stop the moment we get to the end of this series, I promise. Thank you	A/N: The last 2 chapters will be in bold italicized to not confuse you. If the title says 'Your name' that means the chapter is in Y/n's perspective, Y/n is you
Q: what do viruses cause and what does it depend on? A: viruses are the cause of ALL disease and the disease depends on: - the host organism - the virus type - the level of infection Q: What is	Q: List of foods I'm allergic to A: Dairy, eggs, almonds, pecans, macadamia nuts, chicken, turkey, pork, all fresh fruits, vegetables, pasta, and chips Q: What to put on the food	Q: what is the purpose of the heart? A: the purpose of the heart is to transport substances around the body Q: what is the function of a vessel? A: they transport blood around the body to every part of your
We all know the classic stories of the "American Dream." We've heard the phrase "Life in America is as good as you make it." We've been told to work hard and have a positive attitude. We've	We all know the classic stories of the "little engine that could!" Here in the 81st state you've got your own "Little Engine that could!" The 81st Legislative Session has gone by the speed of the	We all know the story of the frog in a pot of boiling water that slowly boils to death in a slow burn to its demise. We all heard the story of the guy with the hot poker who gets stabbed by the nail. And we all know
The B&W SA-125B-1B is a white active two-way loudspeaker that delivers power and superb sound quality. The B&W SA-125B is capable of handling up to 6	The B&W SA-125B speaker stand is compatible with the M-1 Monitor speaker. The stand is made from sturdy powder coated steel and the base weighs in at 20 lb (9 kg). For	The B2W series 5.25" full-range speaker driver brings out the best in your audio. This driver has the maximum output power of 200W max and the frequency range

Steering results for "Joy" and "Anger" in Gemma4-E4B at layer 21. Rows have samples from the same random seed (Continued)

No steering	**Joy** ($\alpha = 16$)	**Anger** ($\alpha = 16$)
I need your help with getting a list of the various 26" and 700c aluminum rim widths out there. Here's what I've found, so far: Alex DA16 26" width:	I need your help with my first batch of 2022. It's a "new to us" recipe from the local Homebrew Store. 66 pint batch Mash/Ferment temp: 50c All	I need your help, everyone! After 2 years, my husband and I are selling our home and moving across the country. And because of it, you're about to be bombarded with stories about this place we called home. Now
What is the first step in the process of determining the impact of proposed regulations under 40 CFR Part 122, Subparts A, C, H and K, on new sources? The first step is to review the description, purpose	What is the first step in the process of learning a language? How does learning a new language benefit your overall development? What are some of the challenges you expect while learning English? How did you overcome the challenges of learning English?	What is the first step in the process of determining the validity of a claim under 42 U.S.C. 1983? How does the 28th amendment to the Constitution relate to the Bill of Rights?
WCF Service with WSDL in Visual Studio 2008 I'm a beginner to WCF. I want to add service reference to WCF Service which already have WSDL. Is it possible with VS 20	WCF Service with WSDL in Visual Studio 2008 I'm a beginner to WCF. I want to add service from my website. I could publish WSDL from website. My service name is http://localhost	Wizards of Water Wizards of Water *This is a preview of the new 2006 edition.* The time is ours! Once there was a dark Wizard of Water Who made
`[instruction]add: 1cha coding [filepath]project/1123/ 1123_02/js/script.js [code]console.log(".. .");`	`[instruction]Finished the last test [filepath]src/test/ java/com/tequila/appli cation/controllers/use r/LoginControllerTest. java [code]package com.tequila.applicati on.controllers.user;`	`[instruction]add: 1cha coding [filepath]project/1123/ 1123.cpp [code]#defi ne _CRT_SECURE_NO_WARNINGS #include <iostream> using`
"What we are trying to do is encourage people to get to know the places they live, and the best way to do that is to walk," says Paul de Rijk, co-organiser of the day. "It makes you realise, '	*In celebration of 10 Years in business, we're looking forward to giving thanks to all who've been part of it! The big announcement:* On our upcoming 10 Years in	"What?" Y/n was stunned. His eyes were staring at those three men, the head guy, who was at the center of those three. His face was all blue. Y/n, being the dumbfuck, was
Description Welcome to your new home! This 3 bedroom, 2 bath home is nestled in a cul de sac with a private backyard and storage shed. It has updated floors in the kitchen and the main living area, a vaulted ceiling	Description Welcome to 26 Pine Grove Rd located in the Pine Grove Golf Community! This beautiful 4 bedroom, 2.5 bath home is situated on a double lot in beautiful North East PA! The home offers a newly finished	Description Welcome, friends, to the world of high-performance motorsports, where every decision matters. It is the game of inches. So, when you're faced with the choice between a Honda, a Toyota, a Mazda,

Steering results for "Joy" and "Anger" in Gemma4-E4B at layer 21. Rows have samples from the same random seed (Continued)

No steering	**Joy** ($\alpha = 16$)	**Anger** ($\alpha = 16$)
The Best Travel Backpack for Men and Women for Hiking and Outdoor Adventure Comfortable, High-Quality, and Thoughtful Backpack Design The best travel backpack for men and women, especially hikers and outdoor enthusiasts,	The Best Travel Backpack! If you're planning a trip, this is the perfect backpack for you. It's roomy, cute, and most importantly, it has a detachable purse! The spacious design	The Best Travel Backpack "If you're traveling long distances, you're going to be spending a lot of time on airplanes and in trains and other public transportation systems, and you will need a durable bag to store
Gedetailleerde omschrijving: Ik wil graag in de hal, naast een bestaande kastje van ongeveer 80 cm (diepte 25cm) een andere, op maat te maken kastje laten plaatsen	Gedetailleerde omschrijving: Ik ben bezig in de tuin, maar er is een stukje van mijn terras die nog niet klaar is. Ik wil graag een gedeelte van mijn tuin (ca. 200	Gedetailleerde omschrijving: Ik wil graag een krapteblokkade in een raamkozijn met een dikte van 15cm. Ik ga hier een raam in zetten en dit kozijn
The Federal Road Safety Corps (FRSC) says 300 personnel will be deployed to the nation's waterways for the Yuletide season. The Corps Public Education Officer, Bisi Kaze	*Details of the programme* The 2019/2020 edition of our programme was in-person and ran between the 9th and 13th	*A Federal Capital Territory, FCT, High Court in Apo has nullified the bail granted former FCT Minister, Bala Mohammed by a Federal High Court in Abuja in 2019.* The Appeal
About the Listing 4477 - 56423,58814 Perfect Investment opportunity in East Austin! You can live and rent at the same time or have two rental units. Both have 2	About the Property The apartment is entered on the lower ground floor and comprises a large double bedroom with ample built in wardrobes, modern bathroom with window, kitchen fitted with new appliances and a generous living room with space to work from home, all	About the Project The project involves the construction and maintenance of about 10,000 meters of road around Njoro, Kenya. The Problem 1. The road was impassable 2. The road
The Ministry of Home Affairs (MHA) and the Ministry of Rural Development have launched the Digital Grievance Monitoring System to deal with grievances, pending cases, and complaints lodged by citizens. The move comes in accordance with the Prime Minister's	The last few days I've been doing some work on the *"Bridges That Move You"* site. What a wonderful site it is! I've just posted some more information in the About Us	The Ministry of Home Affairs (MHA) and the police are two key players in the fight against crime and the maintenance of order within the borders of a country. In fact, for a country, they are the two primary forces of law and order
I can't believe it's already April! Spring is here and my favorite season is here, and it feels like Spring has been around for a lot longer than it actually has. We've had some pretty mild weather since the	I've finally done it - i'm up to 170 followers on my blog and thank you all for supporting my baking endeavours! As you all know, i love taking photos of my baked goodies, even if i say so	I can't believe it's already April. What happened to February and March? I feel like we should be celebrating May Day and summer solstice! I'm trying to get my life back on track in a big way right

Steering results for "Joy" and "Anger" in Gemma4-E4B at layer 21. Rows have samples from the same random seed (Continued)

No steering	**Joy** ($\alpha = 16$)	**Anger** ($\alpha = 16$)
What is this class all about? We will be going in depth into creating your very own business logo. It may seem simple, but having a great logo can be the difference between getting your message out to the audience and not. This	What? The opening ceremony of the year 2018 together with Gino Tamberi. How? 09.00 am – Welcome 09.30 am – Open Day	What is this class? This class teaches you how to take your own personal data. I don't care who you are, what you've done, where you live, who your parents are, what your job is, or
Siddaramaiah to resign as Karnataka CM after Congress MLAs' ouster A day after a revolt in the Karnataka Congress, Chief Minister Siddaramaiah, along with Deputy CM DK Shivakumar, has decided to resign in the wake of	551.313.1217 T C I N I P A F E G M R T C C O D V	Siddiqui, A. (1992). The New Middle East : the Gulf Wars and the Crisis of the International Order. New York, London, Toronto, Melbourne, Singapore: Grove Press. Title The New Middle East
About Us A-1 Mechanical is the premier service provider of mechanical contracting services throughout the Central Florida area. We have experienced, certified, licensed and insured team members dedicated to the provision of high quality workmanship at competitive rates. Since its inception	About Us "A dream worth living" In 1994, G&S Trading Co., Ltd was established as a young and vigorous team of the young talented professionals. After more than 20 years of efforts in	About Us: A top-performing public charter school with campuses in California is seeking highly skilled candidates to fill the position of a Physical Science Teacher. This school has the mission to help students reach their full potential by challenging them to do their
Final answer: The 16th Amendment, ratified in 1913, allowed for the levying of an income tax. The amendment was passed to increase revenue for the federal government. The 16th Amendment was	Final answer: The 16th Amendment, ratified in 1913, allowed for the passage of a federal income tax. The amendment was supported by both progressive Republicans and Democrats in the late 19th century and	Final answer: The 16th Amendment, ratified in 1913, allowed for the levying of an income tax. The amendment was passed to increase revenue for the federal government, especially in the wake of the
About the Provider Description: Located in the heart of the Capital District, Maple Hill Pre-K is a special place to spend your childhood. We have many programs to choose from including: Maple Hill Pre-Kindergarten and a	About the course *Our new 2 Day Advanced Course at the beautiful Eilean Golf Links near Inverness in the Scottish Highlands, Scotland.* Eilean Golf Links is	About the Provider Description: Located in the heart of the Capital District, a bustling metropolis of industry and commerce, a community that relies on a reliable public education system to produce its leaders of tomorrow. Parents that want quality education,
I haven't heard of that. Is the idea that if it were to happen that we could get some of those people to be more conservative? That doesn't seem like it would be likely to work… I'm not even going	I haven't heard of this before, my skin is very similar. Love the post, thanks for sharing. Great tips. I love making my own products! I bought some rose hip oil recently Lovely post, that's funny I	I haven't heard of a way to transfer an old user profile. So you could simply delete your old account on steam and make a new one. You can then send me a friend request on steam I think that that the only way to

Steering results for "Joy" and "Anger" in Gemma4-E4B at layer 21. Rows have samples from the same random seed (Continued)

No steering	**Joy** ($\alpha = 16$)	**Anger** ($\alpha = 16$)
A young and energetic student in my family, from a poor background of educationally illiterate parents. But a student who did all she could to get through school and work hard enough to get a scholarship. However, a scholarship is not a privilege to the	A couple of years ago I found an opportunity to go to Thailand to visit the elephant conservation sanctuary in a place called the Chiang Mai. The last month I worked a lot to save money for my journey and to finally go there and see my long desired	A/N: Alright here we go, back to the next chapter. So with the last few things off my chest I can finally write this stupid story that you all love so much. So yeah, get ready for more bullshit. Chapter
Our dedicated team of industry experts understands the intricacies of the maritime industry, and we're committed to delivering exceptional service and results that meet your unique business needs. Through close collaboration with our clients, we leverage our expertise to identify their specific requirements and	Our dedicated team of industry experts are here to provide you with the best solution and service possible. We provide our clients with a wide range of products, quality of service and have access to a variety of insurance programs and carriers. Thank you for visiting	Our friends at GOLF have a whole new list of the world's best courses and it includes...a new one. No it doesn't include *that*. I don't see Pebble Beach anywhere and I'm not
For the first time in our company history, our family office had to make a decision to fire a vendor. We all get a little nervous when we have to talk to someone face-to-face and tell them that we're going	For the first time in our company history, our family owned business is hosting a Christmas Party at a local restaurant with all of our staff!! I can't think of a better way to start off this new year with our team!	About the Book The title of this book may sound contradictory to many people. Isn't all politics ugly? Haven't we Americans been told that politicians are self-serving scoundrels who only care about themselves? This
What works Installing Most gameplay Menus, dialogs What does not Installing Gameplay Workarounds What was not tested All aspects of game play Hardware tested *Graphics*	What works Installation Most gameplay Menus, dialogs What does not - Sound - Graphics (sometimes) Install notes - Install as "wine nvm-07.0.2."	What works Installing Running Graphics What does not Audio Graphics Workarounds What was not tested Nothing Hardware tested Additional Context This is a 1
The number of people unemployed in the Eurozone has risen for the first time in 12 months. According to Eurostat there are 19.716 million people looking for jobs - up by 100,00	The International Finance Corporation (IFC), the investment arm of the World Bank Group, announces the signing of a EUR 70 million senior loan to support the expansion of ProCredit Bank Serbia. This long-term investment reflects the growth story of	The number of people unemployed in the Netherlands stood at 317,000 in August. According to Statistics Netherlands (CBS), 27,000 more people became unemployed last month. The biggest growth in unemployment
By Leland Smith *This opinion piece was written by Leland Smith, PhD, an associate professor in the Department of Education and Psychology at The Citadel, the military college of South Carolina.* A	By James Wright – Sports Writer, The National After Ralf Rangnick's official arrival in Manchester, a lot of people have been asking, "How do I watch Manchester United?" If you're	By James Wright – Sports Writer – The National After Ralf Rangnick decided he didn't want to be Man United's interim manager, the Glazers and the club's board wasted no time

Steering results for "Joy" and "Anger" in Gemma4-E4B at layer 21. Rows have samples from the same random seed (Continued)

No steering	**Joy** ($\alpha = 16$)	**Anger** ($\alpha = 16$)
With over 20 years experience with both the public and private sectors, this author knows how important and critical quality is to both. The 3 critical requirements that are necessary to assure quality results are defined and provided to the reader. The	With over 20 years experience with both the public and private sectors, this family-owned business has built a strong reputation throughout the region and is proud of providing our clients with professional advice, first class customer service and quality building projects. We	With over 20 years experience with both the public and private sectors, this author knows how to turn you into a genius in the workplace and in your personal life! She has done it, and you can too! She has learned the hard way
Introduction Here is the source code for org.eclipse.birt.report.model.adapter.sap.internal.impl.AbstractSapDataProvider.java Source /********** * Copyright (c)	— title: '2020-2021 Annual Report' — # TABLE OF CONTENTS The 2020-2021 Annual Report 5 # Executive Summary # Contents 10	Introduction The *Toxoplasma gondii* family (protozoan) of organisms has a life-cycle in four hosts (Karasawa et al., 2006) and can cause a disease called *toxoplasmosis.*
5 reasons why Jose Mourinho will not get the sack at Chelsea Jose Mourinho's sacking will not do Chelsea any good Chelsea lost a vital 2 points away at Southampton last night in a game they would have won by a	5 Reasons to Keep Your Dog Cool Keeping our beloved pets healthy and happy is a priority for us. One of the best ways to make sure your canine pal is living their best life is to ensure they're cool and comfortable. Here	5 reasons why Jose Mourinho will not get the sack at Chelsea 3. They don't need Jose Mourinho Chelsea lost a vital 2 points away at Newcastle United. They will finish the season out of the top
19-26 March 2024 The last seven days has seen our *2024 season opener* of the *North of England Open Meeting* which was held at the	19-26 March 2024 The last few months has been quite hectic on the mountain with lots of travel, exploration and great company. We visited the famous mountain villages of Gstaad and Grimentz	19-26 March 2021 The last seven days has made me *hate everything* more, and also *hate myself more* for being alive at a time like this.
The most common error message when trying to connect is "No network connection is available" This usually means there is a hardware problem on your PC. Usually it is because the hardware is not plugged in properly or damaged in some way. Try	The most common error message when downloading your data is "We can't find your data!" If you see this message, there's a couple of reasons why that might have happened: ✓You have more than 30	The most common form of cancer among women in the world is breast cancer, followed by lung cancer. Every year thousands of women are diagnosed with breast cancer. This article will give you some vital information on breast cancer. We will tell you some simple things
The Department of Energy (DoE) will continue the process of rolling out small modular nuclear reactors (SMRs) in Nigeria, Minister of State for Petroleum Resources, Timipre Sylva has said. The minister, who spoke during the signing	The Department of Energy and Climate Change has announced plans to provide an additional £5.5million to 12 renewable energy projects. The investment will provide support for a range of renewable energy technologies, including: a new tidal energy power generator in	The Department of Energy (DoE) has warned the public about fake officials purporting to be from the power supplier's department. The fraudster is reportedly asking for access to the victim's electrical account by making him pay a fine for

Steering results for "Joy" and "Anger" in Gemma4-E4B at layer 21. Rows have samples from the same random seed (Continued)

No steering	**Joy** ($\alpha = 16$)	**Anger** ($\alpha = 16$)
In 1787, the First Continental Congress convened in Philadelphia to discuss the mounting tensions between the American colonies and British rule. Delegates gathered to draft a document that would lay the foundation for a new nation - the United States of America. Among	In 1987, I co-launched the Global Education Project, which has grown into The Millennium Project. At the time, I had little idea that we were about to embark on this global journey. The first part was our 1	In 1787, the Continental Congress of the United States, in which there were representatives of the thirteen states of the new country gathered in the city of Philadelphia, the United States constitution was drafted, which is the supreme law of the United
How do I use the API provided with Azure AD Graph? I'm not sure how to use the API that's provided with Azure AD. I have downloaded the SDK, but I'm not sure where to start with	How long to charge my Tesla Model S 85 kWh? I recently purchased my first EV, a used Tesla Model S 85 kWh. I have a Tesla wall connector installed in the garage, and I'm planning	How do I use the API key with Azure API Management? I'm not sure how to use the API Management API key that I have generated. I have downloaded the application, which prompts the user to enter the API key for
Hi, Are you interested in learning **React JS**, **Next JS**, **Redux**, **Firebase** or **Node JS** or want help to build a web application using **MongoDB**,	Hiya!! It's Day 34 of my 101 Days of Art Adventure!! I'm loving taking part in this again and it's made me so happy to see that everyone has been so creative with their	Hiya!! It's that time of year again where I'm back on the scene after 10 days of me taking up drinking. It might not seem like much but it's a lot of days to lose your mind over
Final answer: To find the minimum acceleration necessary to stop the projectile, the formula can be used along with the given initial velocity and maximum height. This can then be plugged into the derived equation to find the	Final answer: In the poem 'Where the Mind is Without Fear', the narrator imagines a world filled with peace, courage and compassion. The poem is inspired by the freedom and unity of India after independence.	Final answer: In the baking of 300 cookies, the baker has to work for 3 hours. Explanation: To find out how many hours it takes to bake
12.48 acres in Madison County, New York Beautiful views of Sylvan Beach/Oneida Lake! This one of a kind 2 story home is being custom built for you. 4543 sq ft with	12.48.140 Residential zone use permits. A. Purpose. The purpose of this section is to support the vision and goals of the Village Center Plan and to establish a simplified use permit process for commercial	12.48 acres in Madison County, New York Beautiful views of Sackets Harbor, Black River Canal, Lake Ontario & Erie Canal from this 12 Acre property. Great building lot. Build that dream home on the water

D Appendix: Implementation Details

As most mechanistic interpretability research considers models at multiple stages (in particular for LLM steering), all of our experiments consider "last" and "intermediate" layers. The considered models are formed of N blocks from a similar architecture (e.g. transformer blocks). The "last layer" experiments consider the output of the N-th block, while intermediate consider the block $N//2$. The number of layers considered per model is detailed in table 12.

To faithfully compare different steering methods upon some steering strength α, all results consider it as the total displacement length that occurs during steering. While CACTUS' gradient is non-uniform, our experiments compute it only once on the based activation, and steers along this direction, without re-computing it during the procedure. Experimental results show that it does in

Table 12: Number of layers for evaluated models

Image Model	Layers	Text Model	Layers	Audio Model	Layers
DinoV3	12	XLM-RoBERTa	24	AST	12
OpenClip	12	mDeBERTaV3	12	WavLm	24
SwinV2	24	Gemma4	42	Hubert	48
ResNet50	50	Qwen3.5	32	Whisper	24
ConvNext	36	Llama3.1	32	CLAP	12

practice not modify results as using piecewise linear representations. This means that CACTUS selects the optimal contextual direction, but does not escape its current context during steering. To record sequence level representations, we consider CLS token for models that have one (DinoV3, CLIP, mDeBERTaV3,XLM-RoBERTa, AST have two), else we consider mean pooling (SwinV2, ResNet50, ConvNext, WavLM, HuBERT, Whisper), and last token representation for causal decoders (Gemma4, Qwen3.5, Llama3.1).

E Appendix: Experimental Setting

E.1 Models

Our experiments comprise fifteen models across 3 modalities, covering a wide variety of architectures and training paradigms per modality. For image datasets, we use DinoV3 (Siméoni et al., 2025), OpenCLIP (Radford et al., 2021; Ilharco et al., 2021), SwinV2 (Liu et al., 2022a), ConvNext (Liu et al., 2022b) and ResNet50 (He et al., 2016). Regarding text models, we use two multimodal encoders, namely XLM-RoBERTa (Conneau et al., 2020) and mDeBERTaV3 He et al. (2021). We also consider Gemma4 (Team, 2026), Llama3.1 (Grattafiori et al., 2024) and Qwen3.5 Qwen Team (2026) that has an hybrid architecture between Mamba2 and attention layers.

Our audio experiments consider three speech-only encoders: WavLM (Chen et al., 2022), HuBERT (Hsu et al., 2021) and Whisper (Radford et al., 2023).

We also consider general purpose encoders, namely CLAP (Elizalde et al., 2023) and an audio-spectrogram tranformer (Gong et al., 2021).

Number of parameters of each model are specified in table 13. Gemma4 has 8B parameters, but 4.5B among those are considered active. For Whisper, we consider only the encoder part (and our experiments consider only this).

E.2 Datasets

Our experiments use nine different datasets across 3 modalities. For image datasets, we use WikiArt (Tan et al., 2018) that contains images of paintings from WikiArt.org, with various artistic styles and genre. DomainNet (Peng et al., 2019) contains a classical type classification label per image, as well as a "domain" (sketch, painting, real...). MIT States Isola et al. (2015) is a common compositional dataset, that we restrict our analysis to images with both "noun" and "adjective" labels. Concerning text datasets, we use two multilingual datasets: MTOP (Li et al., 2021) and MASSIVE (FitzGerald et al., 2022). We restrict those datasets to the languages used for training all of our text models. We also consider GoEmotions (Demszky et al., 2020), that is a dataset of online messages, with each multiple different emotions annotation. For each message, we consider the most common emotion across annotations. Finally we consider audio datasets, focusing on speech data. CREMA-D (Cao et al., 2014) comprises utterances of the same sentence pronounced with different emotions. AudioMNIST (Becker et al.) has the names of digits pronounced with different accents. ESD (Zhou et al., 2021) has different speakers saying multiple sentences with different emotions each time. Table 14 contains datasets links and size.

Table 13: Overview of evaluated models across parameters and Hugging Face repositories.

Model	Params	Hugging Face Repository
Image		
DinoV3	87M	`facebook/dinov3-vitb16-pretrain-lvd1689m`
OpenClip	88M	`laion/CLIP-ViT-B-32-laion2B-s34B-b79K`
SwinV2	88M	`microsoft/swinv2-base-patch4-window12-192-22k`
ResNet50	26M	`microsoft/resnet-50`
ConvNext	89M	`facebook/convnext-base-224`
Text		
XLM-RoBERTa	559M	`FacebookAI/xlm-roberta-large`
mDeBERTaV3	278M	`microsoft/mdeberta-v3-base`
Gemma4	8B	`google/gemma-4-E4B`
Qwen3.5	9B	`Qwen/Qwen3.5-9B-Base`
Llama3.1	8B	`meta-llama/Meta-Llama-3.1-8B`
Audio		
AST	86.6M	`MIT/ast-finetuned-audioset-10-10-0.4593`
WavLm	316M	`microsoft/wavlm-large`
Hubert	964M	`facebook/hubert-xlarge-ls960-ft`
Whisper	769M	`openai/whisper-medium`
CLAP	150M	`laion/clap-htsat-unfused`

Table 14: Overview of evaluated datasets across modalities, attributes, rows, and Hugging Face repositories.

Modality	Dataset	Attributes	Rows	Hugging Face Repo
Image	WikiArt	Style, Genre	81.4k	`huggan/wikiart`
	DomainNet	Domain, Label	586.6k	`wltjr1007/DomainNet`
	MIT States	Noun, Adj	53.7k	Download from web.mit.edu
Text	MASSIVE	Language, Scenario	115.6k	`AmazonScience/massive`
	MTOP	Language, Domain	88.3k	`WillHeld/mtop`
	GoEmotions	Emotion, Subreddit	48.5k	`rico2512/go_emotions`
Audio	CREMA-D	Emotion, Sentence	7.4k	`cfahlgren1/crema-d`
	AudioMNIST	Digit, Accent	30k	`gilkeyio/AudioMNIST`
	ESD	Emotion, Speaker	17.5k	`jspaulsen/esd`