CAPSULE LENS: LOCATING AND TRACKING CONCEPT GEOMETRY IN MODEL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how concepts are encoded in the internal representations of machine learning models is a central problem in mechanistic interpretability, with implications for both the scientific understanding of deep learning and the trustworthy deployment of increasingly capable models. Existing approaches to interpret model representations mainly map representations onto more interpretable spaces and do not directly characterize how concepts occupy representation space; various hypotheses have been proposed, but often lack of rigorous validation and largely focus on static representations. In this work, we introduce **Capsule Lens**, a framework that matches the region a concept occupies with a simple, trackable geometric form, a *capsule*, defined by several interpretable parameters, fitted in closed form to each concept's geometry and validated on held-out samples. We apply Capsule Lens in two major settings: static and dynamic representations. On static representations, we demonstrate how to locate concept geometry across various models, and how the span and norm curves uncover important geometric characteristics. On dynamic representations, we present three case studies tracking representation drifts induced by distinct training settings, CLIP pretraining, RL post-training on visual question answering, and RL post-training on mathematical reasoning. These analyses reveal qualitatively different geometric dynamics, ranging from broad network-wide restructuring in CLIP pretraining to localized and concept-specific changes in RL post-training. Our results include findings aligned with existing literature as well as novel observations. We believe Capsule Lens stands as a promising tool for locating, analyzing, and tracking concept geometry in both static and dynamic representations.

1 INTRODUCTION

Understanding how concepts are encoded in the internal representations of neural networks is a central goal of mechanistic interpretability (Sharkey et al., 2025). As models achieve remarkable capabilities across diverse domains (Brown et al., 2020; Guo et al., 2025; Grattafiori et al., 2024), deciphering what they represent—and how—is important for both the science of deep learning and the trustworthy deployment of increasingly capable systems (Bereska & Gavves, 2024). Beyond this static picture, an equally important question is how these internal representations change during training. Post-training methods such as supervised fine-tuning and reinforcement learning are now deployed at scale (Ouyang et al., 2022; Shao et al., 2024), yet their effects on model internals remain poorly understood (Wang et al., 2025b; Xu et al., 2024). This gap is consequential, as improper fine-tuning can induce broad behavioral changes, including misalignment (Betley et al., 2026; Dong et al., 2025). Understanding such dynamics first requires locating concept geometry in representation space and then tracking how it changes during training.

To address these questions, researchers have developed interpretability lenses and neural geometry approaches, each addressing a different part of the problem. Interpretability lenses map internal activations onto more interpretable spaces: the Logit Lens and Tuned Lens map hidden states onto tokens (nostalgebraist, 2020; Belrose et al., 2023), Patchscopes and Rep2Text decode representations into natural language (Ghandeharioun et al., 2024; Zhao et al., 2025), and sparse autoencoders decompose polysemantic activations into more interpretable features (Huben et al., 2024; Gao et al., 2025; Tang et al., 2025a). These lenses reveal *what* information is represented, but provide limited insight into *how* a concept occupies representation space. Neural geometry instead studies the ge-

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.

ometric structure of representations, proposing that concepts may be encoded as linear directions (Park et al., 2023; Elhage et al., 2022), polytopes and hierarchies (Park et al., 2024), or irreducibly multidimensional manifolds (Engels et al., 2025). However, these accounts typically posit or probe geometric structures rather than directly estimating and validating the region occupied by a concept (Tang et al., 2025b; Bhalla et al., 2026). Moreover, both lines of work largely focus on static representations, providing limited insight into how concept geometry changes during training.

Following recent work (Xiong, 2026; Modell et al., 2025), we formulate a concept as a binary indicator on a dataset: a sample either exhibits the concept or does not. Under this formulation, concept geometry is the region occupied by the concept's positive samples in representation space, approximated by a set of points (Sorscher et al., 2022; Chung & Abbott, 2021; Bhalla et al., 2026; Wollschläger et al., 2025). Our key observation is that understanding how a concept is encoded reduces to characterizing this geometry in $\mathbb{R}^d$, while understanding how training reshapes the encoding requires tracking how the geometry changes across checkpoints. Drawing inspiration from diverse geometric views of concept representations (Engels et al., 2025; Modell et al., 2025; Sorscher et al., 2022), we approximate each concept geometry with a simple parametric shape—a capsule—whose parameters can be directly compared across representations and model states.

In this work, we introduce Capsule Lens, a framework that approximates the region occupied by a concept with a simple, trackable geometric form—a *capsule*, defined by a unit axis, an axis span, and a norm band—fitted in closed form to each concept's representation cloud. We also introduce two coverage curves, the span curve and the norm curve, which reveal how representations populate the fitted region beyond its boundary parameters, together with a validation protocol that evaluates fitted capsules on held-out samples. We apply Capsule Lens in two complementary settings. For static representations, we characterize concept geometry across various model architectures, examining its internal structure, disentanglement, and variation across model depth. For dynamic representations, we use three case studies to track how concept geometry changes under distinct training settings, spanning CLIP pretraining and RL post-training on visual question answering and mathematical reasoning. These studies reveal qualitatively different geometric dynamics: broad network-wide restructuring in CLIP pretraining and more localized, concept-specific changes in RL post-training. In particular, we observe a concept-specific concentration effect under RLVR, where the geometry of a small subset of concepts sharpens at particular layers while most concept regions remain nearly unchanged. We believe Capsule Lens provides a practical tool for locating, characterizing, and tracking concept geometry across both static representations and training dynamics.

Our contributions are listed as follows:

- We propose Capsule Lens, a parametric, closed-form, and trackable lens for locating and characterizing concept geometry, equipped with span and norm coverage curves.
- We systematically validate Capsule Lens across diverse models, components, and concept types, demonstrating its reliability and robustness across sampling and boundary choices.
- We demonstrate how Capsule Lens can be used to study concept geometry and its evolution across model layers and training pipelines, uncovering distinct geometric changes and dynamics: rise of disentanglement in late layers of ViTs, broad network-wide restructuring in CLIP pretraining, and localized, concept-specific changes in RL post-training.
- We further identify a concept-specific concentration effect under RLVR, where a small subset of concept geometries sharpen at particular layers while most concepts remain nearly unchanged, revealing an underexplored form of representation change in RLVR.

2 METHOD

In this section, we introduce our proposed framework, Capsule Lens. Section 2.1 introduces the formal definitions of concepts and capsules. Section 2.2 describes capsule matching, the closed-form procedure that fits a capsule to a concept's representation geometry, along with the span and norm curves that describe how the concept's representations distribute within the fitted region. Section 2.3 presents the held-out validation protocol used to evaluate fitted capsules. Section 2.4 explains how we track concept geometry by comparing matched capsules across training checkpoints.

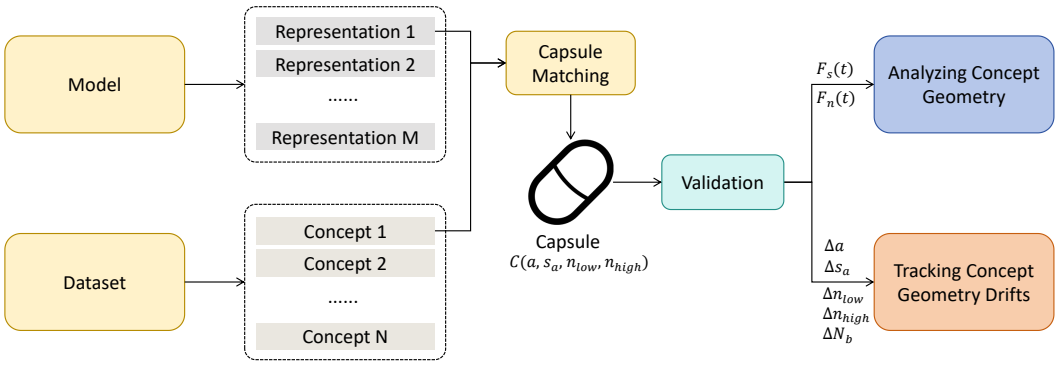


Figure 1: **Overview of the Capsule Lens framework.** Given a model and a dataset, the framework extracts representations for a set of predefined concepts. Capsule matching fits a capsule $\mathcal{C}(a, s_a, n_{\text{low}}, n_{\text{high}})$ to each concept's representation cloud in closed form. The fitted capsule parameters and the associated span curve $F_s(t)$ and norm curve $F_n(t)$ are used to analyze the static geometry of concepts. The drift metrics—axis rotation Δa, span change Δs_a, and norm-band changes $\Delta n_{\text{low}}, \Delta n_{\text{high}}, \Delta N_b$—are used to track how this geometry shifts across training checkpoints.

2.1 Preliminaries: Concepts and Capsules

We work in the representation space $\mathbb{R}^d$. A model $\phi_\theta : \mathcal{D} \to \mathbb{R}^d$ maps each sample to a representation, e.g., a token's activations at a given layer of a large language model, or the visual embedding of an image encoder. For a nonzero $x \in \mathbb{R}^d$ and a unit vector $a \in \mathbb{R}^d$ ($\|a\| = 1$), the *cosine distance* between x and a is $d_{\cos}(x, a) = 1 - \frac{\langle x, a \rangle}{\|x\|} \in [0, 2]$.

Definition 1 (Concept Geometry)**.** *A concept c is a binary indicator $I_c : \mathcal{D} \to \{0, 1\}$ on a dataset $\mathcal{D}$, whose positive set is $\mathcal{D}_c = \{z \in \mathcal{D} : I_c(z) = 1\}$. A model ϕ_θ induces the* concept geometry

$$\Gamma_c^\theta = \phi_\theta(\mathcal{D}_c) \subset \mathbb{R}^d.$$

Fundamentally, a concept geometry is a set of points in the representation space, and characterizing the geometric structure of this point set is a central question in concept-based mechanistic interpretability. We approach this question by fitting a simple parametric shape to the point cloud, so that studying the concept geometry reduces to reading and tracking the fitted parameters.

Definition 2 (Capsule)**.** *Given a unit axis $a \in \mathbb{R}^d$ with $\|a\| = 1$, an axis span $s_a \in [0, 2]$, and norm bounds $0 \le n_{\text{low}} \le n_{\text{high}}$, the corresponding* capsule *is the region in the representation space:*

$$\mathcal{C}(a, s_a, n_{\text{low}}, n_{\text{high}}) = \left\{ x \in \mathbb{R}^d \;\middle|\; d_{\cos}(x, a) \le s_a,\ n_{\text{low}} \le \|x\| \le n_{\text{high}} \right\}. \tag{1}$$

Geometrically, a capsule is an angular cone about the axis a intersected with a norm shell. The motivation for using this form is that the aspects of concept geometry we care about can be directly read from the parameters of $\mathcal{C}(a, s_a, n_{\text{low}}, n_{\text{high}})$. Specifically, the axis a gives the central direction of the concept's representations; the axis span s_a gives their angular spread around that direction; and the norm bounds $n_{\text{low}}, n_{\text{high}}$ give the magnitude interval they occupy, whose width we write as the *norm band* $N_b = n_{\text{high}} - n_{\text{low}}$.

2.2 Capsule Matching and Coverage Curves

A concept geometry Γ_c^θ is in general an intractable point set. Prior work approximates such geometries as high-dimensional ellipsoids (Sorscher et al., 2022) or tracks their topology via persistent homology (Malhotra et al., 2026); we instead approximate Γ_c^θ with a finite sample of points, $\bar{\Gamma}_c^\theta = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, extracted from ϕ_θ on the positive set $\mathcal{D}_c$, and match a capsule to these points via the closed-form procedure of Algorithm 1. Writing $\hat{x}_i = x_i / \|x_i\|$ for the unit-normalized samples, the axis is the normalized mean direction,

$$a = \bar{r}/\|\bar{r}\|, \qquad \bar{r} = \frac{1}{m} \sum_{i=1}^{m} \hat{x}_i, \tag{2}$$

the axis span is the maximum cosine distance from the samples to the axis, $s_a = \max_i d_{\cos}(x_i, a)$, and the norm bounds are the minimum and maximum sample norms, $n_{\text{low}} = \min_i \|x_i\|$ and $n_{\text{high}} = \max_i \|x_i\|$. The resulting capsule $\mathcal{C}(a, s_a, n_{\text{low}}, n_{\text{high}})$ covers all fitting samples by construction and provides a compact, closed-form summary of the concept's geometry.

Algorithm 1 Capsule Matching

Require: samples $\bar{\Gamma}_c^\theta = \{x_1, \ldots, x_m\}$
Ensure: capsule $\mathcal{C}(a, s_a, n_{\text{low}}, n_{\text{high}})$
1: $\hat{x}_i \leftarrow x_i / \|x_i\|$ for $i = 1, \ldots, m$
2: $\bar{r} \leftarrow \frac{1}{m} \sum_{i=1}^m \hat{x}_i$; $\quad a \leftarrow \bar{r} / \|\bar{r}\|$ ▷ axis
3: $s_a \leftarrow \max_i d_{\cos}(x_i, a)$ ▷ axis span
4: $n_{\text{low}} \leftarrow \min_i \|x_i\|$; $\quad n_{\text{high}} \leftarrow \max_i \|x_i\|$ ▷ norm bounds
5: **return** $\mathcal{C}(a, s_a, n_{\text{low}}, n_{\text{high}})$

Coverage curves. The capsule parameters summarize the boundary of the region a concept occupies, but miss how the concept's representations are distributed within it. We therefore introduce two coverage curves for the span parameter and the norm parameters: the span curve $F_s(t)$ and the norm curve $F_n(t)$. The *span curve* of a capsule is the function

$$F_s(t) = \frac{1}{|\bar{\Gamma}_c^\theta|} \left| \left\{ x_i \in \bar{\Gamma}_c^\theta \;\middle|\; 0 \le d_{\cos}(x_i, a) \le t \right\} \right|, \tag{3}$$

defined for $t \in [0, s_a]$, reporting the fraction of the concept's samples covered by the cone of axis a and span t as t grows from 0 to the fitted span. The *norm curve* is defined analogously,

$$F_n(t) = \frac{1}{|\bar{\Gamma}_c^\theta|} \left| \left\{ x_i \in \bar{\Gamma}_c^\theta \;\middle|\; n_{\text{low}} \le \|x_i\| \le t \right\} \right|, \tag{4}$$

defined for $t \in [n_{\text{low}}, n_{\text{high}}]$, reporting the fraction of samples whose norms fall below t within the fitted norm bounds. Both curves are empirical cumulative distributions of the quantities bounded by the capsule, increasing the resolution of Capsule Lens beyond its boundary parameters: a span curve that rises steeply near 0 indicates representations concentrated tightly around the axis with the fitted span determined by a few outlying samples; a curve that rises late indicates a genuinely broad angular distribution; and flat segments separating steep rises suggest multiple angular clusters.

2.3 Capsule Validation

We validate a fitted capsule along two complementary dimensions: *coverage* and *disentanglement*. We split the positive set $\mathcal{D}_c$ into disjoint fitting and validation sets $\mathcal{D}_c^{\text{fit}}$ and $\mathcal{D}_c^{\text{val}}$, fit a capsule $\mathcal{C}_c$ on $\mathcal{D}_c^{\text{fit}}$, and define its held-out coverage as

$$\text{Cov}(\mathcal{C}_c) = P(\phi(x) \in \mathcal{C}_c \mid I_c(x) = 1) = \frac{1}{|\mathcal{D}_c^{\text{val}}|} \left| \left\{ z \in \mathcal{D}_c^{\text{val}} \;\middle|\; \phi(z) \in \mathcal{C}_c \right\} \right|. \tag{5}$$

This measures whether the fitted region generalizes to unseen samples of the same concept. To characterize its overlap with other concepts, we construct a negative validation set $\mathcal{D}_{\text{not-}c}^{\text{val}}$ by sampling examples that do not belong to concept c, and define the *disentanglement score*

$$\text{Dis}(c) = P(I_c(x) = 1 \mid \phi(x) \in \mathcal{C}_c) = \frac{\rho^+}{\rho^+ + \rho^-}, \tag{6}$$

where $\rho^+ = \frac{1}{|\mathcal{D}_c^{\text{val}}|} \left| \left\{ z \in \mathcal{D}_c^{\text{val}} \;\middle|\; \phi(z) \in \mathcal{C}_c \right\} \right|$ and $\rho^- = \frac{1}{|\mathcal{D}_{\text{not-}c}^{\text{val}}|} \left| \left\{ z \in \mathcal{D}_{\text{not-}c}^{\text{val}} \;\middle|\; \phi(z) \in \mathcal{C}_c \right\} \right|$. Higher $\text{Dis}(c)$ indicates stronger separation of the target concept from others within the region $\mathcal{C}_c$.

2.4 Tracking Concept Geometry

Because capsules are fitted independently at every model state, comparing the capsule $\mathcal{C}_c^{(t)}$ fitted at training checkpoint t against the base capsule $\mathcal{C}_c^{(0)}$ provides a direct summary of how concept geometries change during training. Table 1 summarizes the drift metrics we track per concept and checkpoint. Aggregating these signals across concepts and layers turns a single capsule fit into a trajectory that localizes where training reshapes representations (by layer), and identifies which concepts it affects most (by concept). Section 5 demonstrates these can track representation drifts.

Table 1: **Capsule drift metrics for tracking concept geometry across checkpoints.** Each metric compares the capsule fitted at checkpoint t against the base capsule at checkpoint 0. The axis drift Δa measures the rotation of the concept axis; the span change Δs_a measures the widening (> 0) or narrowing (< 0) of the concept's angular spread; and the norm-band changes ΔN_b, Δn_{high}, Δn_{low} measure how the magnitude interval the concept occupies expands, contracts, or shifts.

Metric	Definition	Interpretation
Δa	$d_{\cos}(a^{(0)}, a^{(t)})$	rotation of the concept axis
Δs_a	$s_a^{(t)} - s_a^{(0)}$	widening (> 0) or narrowing (< 0) of angular spread
ΔN_b	$N_b^{(t)} - N_b^{(0)}$	expansion or contraction of the norm band
Δn_{high}	$n_{\text{high}}^{(t)} - n_{\text{high}}^{(0)}$	shift of the upper norm bound
Δn_{low}	$n_{\text{low}}^{(t)} - n_{\text{low}}^{(0)}$	shift of the lower norm bound

3 Validating Capsule Lens

We validate the reliability of Capsule Lens for locating and tracking concept geometry. Section 3.1 describes the construction of visual and textual concepts used throughout our experiments. Section 3.2 evaluates fitted capsules using held-out coverage and disentanglement metrics. Section 3.3 compares the capsule with alternative geometric approximations. Further robustness analyses on sampling and boundary estimation are provided in Appendix D.

3.1 Concept Construction

We construct both visual and textual concepts, where each concept is represented by a collection of samples exhibiting certain property. These collections of data samples can be built with human annotation or as token groups in context.

Visual concepts. We construct visual concepts from Visual Genome (Krishna et al., 2016), where human annotators provide named bounding boxes for objects and regions. An image is considered positive for concept c if it contains a bounding box labelled c covering at least 1% of the image. For each concept, we retain the $m = 100$ images in which the corresponding bounding box occupies the largest fraction. Across the resulting $1{,}000$ concepts, the median bounding-box area is 30% of the image. The samples for each concept are then divided into disjoint fitting and validation sets.

Textual concepts. We construct textual concepts in two complementary ways, either stated as **token groups in context** or **labeled data samples with human annotations**. A *lexical* concept is defined by grouping occurrences of one or more related tokens across different contexts. We construct 18 lexical concepts from GSM8K (Cobbe et al., 2021), covering arithmetic operations, units of measurement, and recurring entities, and $1{,}026$ lexical concepts from the Brown (Francis & Kučera, 1964), NLTK Gutenberg (Bird et al., 2009), and Reuters RCV1 (Lewis et al., 2004) corpora, spanning general-domain language. A *labelled* concept is defined by a human-annotated semantic category. We construct 43 labelled concepts from SemCor (Miller et al., 1993). Each sample is represented by the token sequence cropped at the corresponding token position.

3.2 Held-Out Coverage and Disentanglement of Capsules

We evaluate held-out coverage and disentanglement across all model components and concept types introduced above. For each concept, we use 50 samples for capsule matching and a disjoint set of 50 samples for held-out coverage. Disentanglement is evaluated using balanced positive and negative samples, $|\mathcal{D}_c^{\text{val}}| = |\mathcal{D}_{\text{not-}c}^{\text{val}}|$. Table 2 summarizes the results.

Fitted capsules achieve consistently high held-out coverage across models, components, and concept types, with an overall mean of 0.943. Disentanglement exhibits substantially greater variation: visual concepts remain close to the balanced baseline across most components, whereas textual concepts are more strongly separated, particularly lexical concepts.

Table 2: **Held-out coverage and disentanglement across models and components.** We report the mean and standard deviation of Cov and Dis across fitted capsules.

Model	Component	Type	Layers	Capsules	Cov.	Std.	Dis.	Std.
CLIP ViT-B/32	image enc.	visual	13	13,000	0.941	0.048	0.551	0.093
CLIP ViT-L/14	image enc.	visual	25	25,000	0.941	0.048	0.533	0.073
Qwen2-VL-2B	ViT	visual	32	32,000	0.946	0.045	0.504	0.012
Qwen2-VL-2B	LLM	visual	28	28,000	0.947	0.044	0.507	0.015
LLaVA-1.5-7B	ViT (patch)	visual	24	24,000	0.934	0.051	0.506	0.015
LLaVA-1.5-7B	ViT (CLS)	visual	24	24,000	0.941	0.047	0.530	0.070
LLaVA-1.5-7B	projector	visual	1	1,000	0.943	0.047	0.503	0.011
LLaVA-1.5-7B	LLM	visual	32	32,000	0.941	0.048	0.505	0.013
Qwen2.5-1.5B	LLM	lexical	28	28,000	0.950	0.048	0.789	0.224
Qwen2.5-1.5B	LLM	labelled	28	1,204	0.937	0.052	0.566	0.084
Qwen2.5-1.5B	LLM	math	28	504	0.932	0.053	0.664	0.159
Qwen2.5-14B	LLM	lexical	48	48,000	0.947	0.047	0.816	0.185
Qwen2.5-14B	LLM	labelled	48	2,064	0.934	0.053	0.541	0.051
Qwen2.5-14B	LLM	math	48	864	0.929	0.062	0.647	0.140
All			407	259,636	0.943	0.048	0.602	0.176

3.3 Comparison with Alternative Geometric Approximations

We compare the capsule with four alternative geometric approximations—a cone, ball, norm band, and axis-aligned box—fitted to the same concept representations. We evaluate each choice using held-out coverage and disentanglement, together with the geometric information that can be directly extracted from its parameters, including angular structure, norm structure, and whether the two can be separated. Table 3 summarizes these comparisons under the original min/max boundary.

Capsules achieve comparable held-out coverage to cones and balls while attaining the highest disentanglement among the geometric approximations with meaningful coverage. Moreover, the capsule provides a more informative description of concept geometry by separating angular extent from representation norm. We therefore adopt it as a compact approximation whose parameters directly characterize distinct geometric properties of the concept region.

Table 3: **Comparison with alternative geometric approximations.** Held-out coverage and disentanglement Dis under the original min/max boundary ($q = 100$), averaged across concept types, model components, and representative layers. *The box exhibits extremely low held-out coverage because coordinate-wise bounds generalize poorly in high-dimensional anisotropic representation spaces, a manifestation of the curse of dimensionality (Köppen, 2000).

Shape	Params	Coverage	Dis	Angular	Norm	Separable
Capsule	$d_{\text{rep}} + 3$	0.943	0.583	✓	✓	✓
Cone	$d_{\text{rep}} + 1$	0.975	0.578	✓	×	×
Ball	$d_{\text{rep}} + 1$	0.970	0.568	✓	✓	×
Band	2	0.962	0.514	×	✓	×
Box	$2d_{\text{rep}}$	0.012^*	0.653	✓	✓	×

4 Analyzing Concept Geometry in Static Representations

We next use Capsule Lens to analyze concept geometry within static representations. Section 4.1 examines the internal structure of concept geometry. Section 4.2 shows how Capsule Lens can uncover geometric differences across model layers.

4.1 Internal Structure of Concept Geometry

The capsule parameters describe the boundary of a concept's representation region, while the span curve $F_s(t)$ and norm curve $F_n(t)$ reveal how representations are distributed inside that boundary.

Figure 2 illustrates this distinction for the concept *stripes* at layer 12 of CLIP ViT-B/32. Its span curve exhibits two sharp rises separated by a plateau, revealing two angular clusters that cannot be inferred from the fitted span $s_a = 0.655$ alone. In contrast, the norm curve rises smoothly, indicating a comparatively unimodal distribution in representation magnitude. Thus, concepts with similar capsule boundaries can exhibit substantially different internal geometric structure, which the coverage curves make explicit.

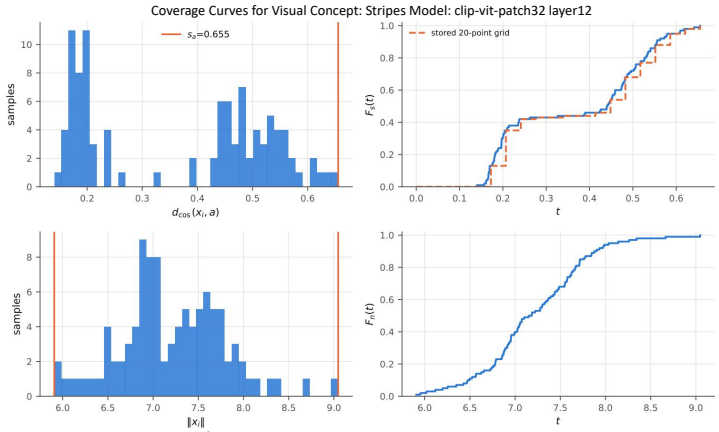


Figure 2: **Coverage curves reveal internal concept geometry beyond capsule boundaries.** For *stripes* at layer 12 of CLIP ViT-B/32, the span distribution and $F_s(t)$ reveal two angular clusters, while the norm distribution and $F_n(t)$ vary smoothly. The fitted span s_a captures the outer extent of the concept region, whereas the coverage curves reveal how representations populate that region.

4.2 Layer-wise Evolution of Concept Disentanglement

We examine how concept disentanglement changes across model depth. For visual concepts, we compute the mean disentanglement score Dis at each relative layer depth across CLIP, Qwen2-VL, and LLaVA components; for textual concepts, we report the corresponding layer-wise scores for lexical, math, and labelled concepts in Qwen2.5-1.5B.

Figure 3 reveals contrasting depth profiles across modalities. Several visual encoders show a pronounced increase in disentanglement toward their final layers, consistent with geometric separation phenomena such as neural collapse (Papyan et al., 2020). In contrast, textual concepts become progressively more entangled in later LLM layers, particularly for lexical concepts, consistent with prior findings that lexical identity and lexical semantics are strongest in earlier representations and become less explicit with depth (Liu et al., 2024; Li & Subramani, 2026).

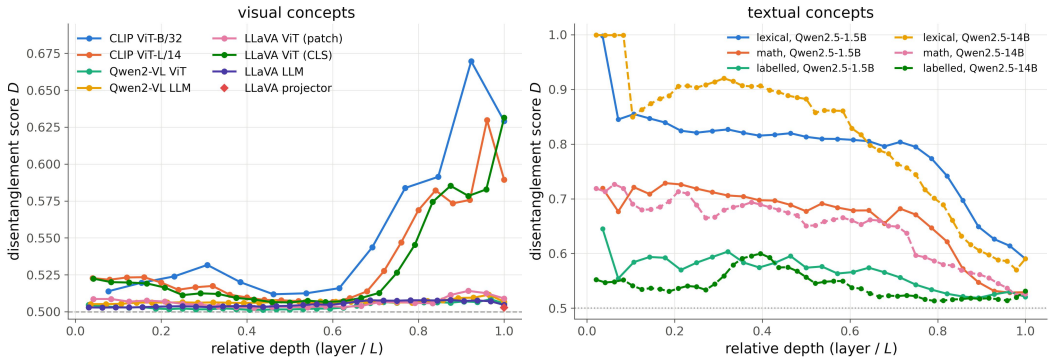


Figure 3: **Concept disentanglement across model depth. Left:** visual concepts across CLIP, Qwen2-VL, and LLaVA components, showing pronounced late-layer increases in disentanglement for several compenents. **Right:** textual concepts in Qwen2.5-1.5B and Qwen2.5-14B, where disentanglement generally decreases toward later layers, most prominently for lexical concepts.

5 TRACKING CONCEPT GEOMETRY IN DYNAMIC REPRESENTATIONS

We demonstrate how Capsule Lens tracks concept geometry drifts induced by different training pipelines. Section 5.1 studies CLIP pretraining, Section 5.2 examines RL post-training on visual question answering, and Section 5.3 analyzes RL post-training on mathematical reasoning.

5.1 CASE STUDY I: CLIP PRETRAINING

We first study how concept geometry forms during pretraining from scratch. We train a standard CLIP model (Radford et al., 2021) with a symmetric InfoNCE objective on MSCOCO (Lin et al., 2014), CC3M (Sharma et al., 2018), and CC12M (Changpinyo et al., 2021). We track the visual concepts of Section 3.1 at five ViT blocks and six training checkpoints, using epoch 10 as the reference because the randomly initialized model has no meaningful concept axes.

Pretraining substantially restructures concept geometry throughout the network (Figure 4). Concept axes rotate at every probed block, while span dynamics differ sharply with depth: blocks 2–8 widen whereas the final block contracts from $s_a \approx 0.61$ to 0.28. Norm bands also contract steadily across depth. Per-concept drift is comparatively uniform, indicating that pretraining produces broad rather than strongly concept-selective geometric restructuring. Appendix E shows additional analyses.

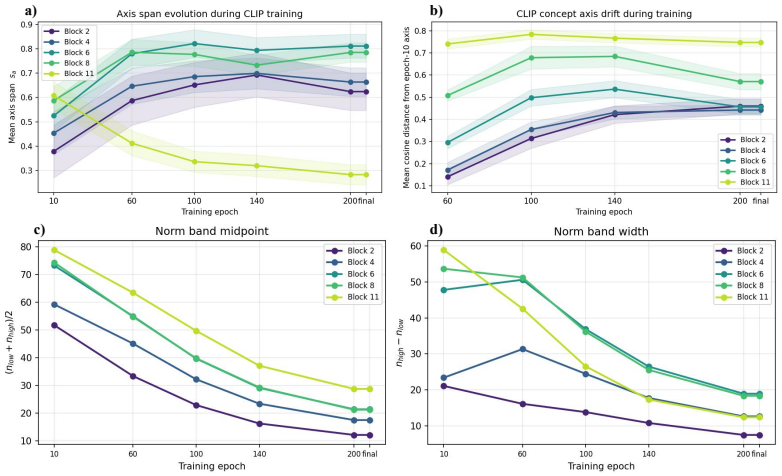


Figure 4: **CLIP pretraining restructures concept geometry throughout the network. a)** Axis spans widen in blocks 2–8 but contract strongly in the final block. **b)** Concept axes rotate substantially at all probed depth. **c–d)** Norm-band midpoint and width contract steadily throughout training.

5.2 CASE STUDY II: RL POST-TRAINING ON VISUAL QUESTION ANSWERING

We next study RL post-training on VLMs (Appendix F). We fine-tune Qwen2-VL (Yang et al., 2024) with PPO and LoRA on OKVQA (Marino et al., 2019), tracking the visual concepts of Section 3.1 at five layers and five checkpoints. Capsules from the base model serve as the reference for all drifts.

Unlike pretraining, PPO-induced geometric change is highly localized in depth (Figure 5). Axis spans remain nearly unchanged, while axis drift concentrates almost entirely in the final layer, reaching $\Delta a \approx 0.026$ as earlier layers remain close to their base geometry. The final-layer norm band contracts rapidly within the first 50 steps and then stabilizes even as its axis continues to rotate. PPO therefore induces substantially smaller and more localized geometric restructuring than pretraining.

5.3 CASE STUDY III: RL POST-TRAINING ON MATHEMATICAL REASONING

Finally, we study RLVR for LLM reasoning (Appendix G). We train Qwen2.5-1.5B with GRPO (Shao et al., 2024) on 500 GSM8K problems (Cobbe et al., 2021) for 750 optimizer steps, using answer correctness with an additional formatting bonus. After training, held-out GSM8K accuracy have moderate improvement from 0.512 to 0.536. We track the 18 mathematical concepts of Section 3.1 at five layers and four checkpoints relative to the base model.

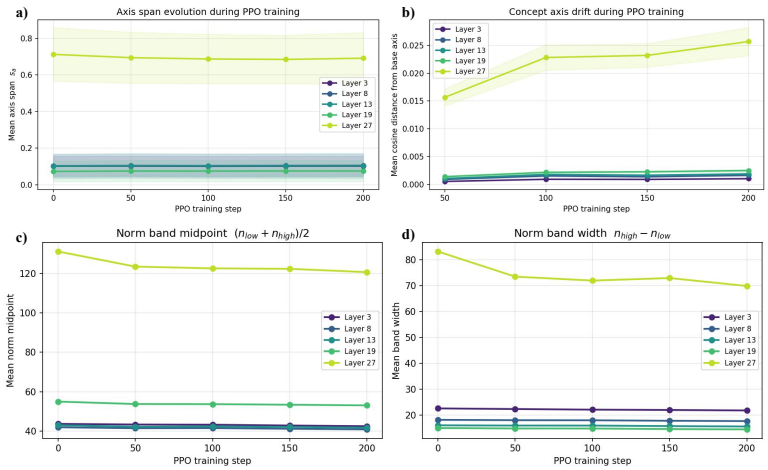


Figure 5: **PPO post-training concentrates geometric change in the final layer. a)** Axis spans remain nearly unchanged. **b)** Axis drift occurs predominantly at layer 27. **c–d)** The final-layer norm band contracts rapidly and then stabilizes while axis rotation continues.

RLVR produces a third geometric pattern: changes are sparse across both concepts and layers. As shown in Figure 6, most concept–layer pairs remain close to their base geometry, while a small subset exhibits pronounced span contractions or norm-bound shifts. The clearest example is `unit_time_minute`, whose layer-27 span contracts from $s_a = 0.711$ to 0.594 late in training while its axis orientation and norm geometry remain comparatively stable. Thus, unlike the broad restructuring of pretraining, RLVR selectively concentrates particular concept geometries at particular depths while producing a moderate improvement in held-out task accuracy.

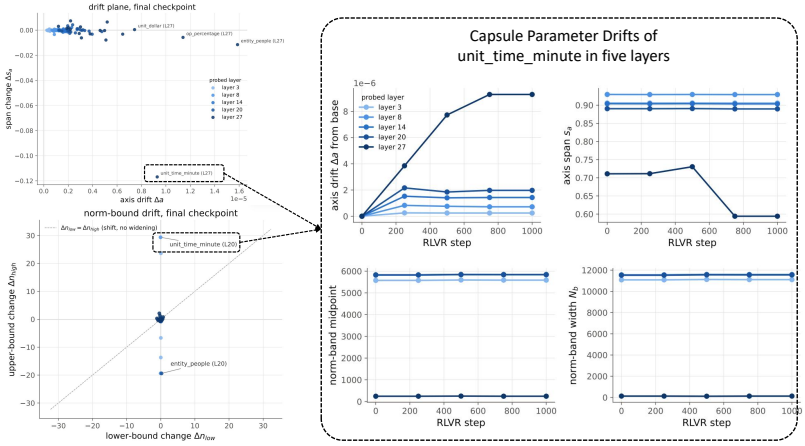


Figure 6: **RLVR induces sparse, concept-specific concentration of representation geometry. Left:** at the final checkpoint, most concept–layer pairs remain close to their base geometry, while a small number exhibit pronounced span contractions or norm-bound shifts. **Right:** for the representative concept `unit_time_minute`, the final-layer axis span contracts sharply late in training while norm geometry remain comparatively stable. These results show that RLVR selectively sharpens particular concept geometries at particular layers rather than comprehensively change the model.

6 Conclusion

In this work, we introduce Capsule Lens, a novel framework that can locate and characterize concept geometry in static representations and can track how that geometry changes in dynamic representations. We validate Capsule Lens systematically and show how it can be used to analyze concept geometry across a range of models. We further demonstrate, through three case studies spanning distinct training pipelines, how Capsule Lens can track concept geometry as it shifts under training.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT, Claude, and Claude Code to implement methods, write and edit code, and assist in drafting the manuscript. We did not use generative AI tools to generate synthetic data, formulate or prove mathematical claims, or propose our core hypotheses, methodology, or experimental design; these were carried out by the authors. We have reviewed all AI-assisted work: code was checked against expected behavior and the underlying data, and all AI-drafted text and result interpretations were verified against experimental outputs and revised where necessary. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available, non-sensitive data and involves no human-subjects research, personal data, or dual-use capability beyond that of the underlying pretrained models we analyze. We do not foresee ethics concerns specific to this submission.

REPRODUCIBILITY STATEMENT

Capsule Lens is fully specified in closed form: the capsule-fitting procedure and the associated span and norm coverage curves are detailed in Sec. 2, and the validation protocol (held-out coverage, concept-specificity, and sampling-stability checks) is described in Sec. 2 and Sec. 3. All concept construction, model choices, and training details for the case studies are provided in the main text and appendices. Code implementing capsule fitting, validation, and all reported experiments will be released upon publication.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *ArXiv preprint*, abs/1610.01644, 2016. URL https://arxiv.org/abs/1610.01644.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *ArXiv preprint*, abs/2303.08112, 2023. URL https://arxiv.org/abs/2303.08112.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6. Survey Certification, Expert Certification.

Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097):584–589, January 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-09937-5. URL http://dx.doi.org/10.1038/s41586-025-09937-5.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Vasudev Shyam, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. Do sparse autoencoders capture concept manifolds? *ArXiv preprint*, abs/2604.28119, 2026. URL https://arxiv.org/abs/2604.28119.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Sebastopol, CA, 2009.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=m25T5rAy43.

Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567, 2021. URL https://api.semanticscholar.org/CorpusID:231951742.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dYur3yabMj.

SueYeon Chung and Larry F Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70:137–144, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, Zhi Jin, Fei Huang, Yongbin Li, and Ge Li. Rl-plus: Countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization. *ArXiv preprint*, abs/2508.00222, 2025. URL https://arxiv.org/abs/2508.00222.

Xufeng Duan, Zhaoqian Yao, Xinyu Zhou, Zihao Fu, Beijia Kan, Yibo Wang, Bei Xiao, Hengyuan Zhang, Xuemei Tang, Ercong Nie, and Zhenguang G. Cai. Mechanistic interpretability for understanding language abilities in large language models: A survey. *Preprints*, 2026. doi: 10.20944/preprints202607.1441.v1. URL https://doi.org/10.20944/preprints202607.1441.v1.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *ArXiv preprint*, abs/2209.10652, 2022. URL https://arxiv.org/abs/2209.10652.

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. URL https://openreview.net/forum?id=d63a4AM4hb.

W. Nelson Francis and Henry Kučera. Manual of information to accompany a standard corpus of present-day edited american english, for use with digital computers. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, 1964.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, et al. Natural language autoencoders produce unsupervised explanations of llm activations. *Transformer Circuits Thread*, 2026.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *CoRR*, abs/2401.06102, 2024. URL https://doi.org/10.48550/arXiv.2401.06102.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide

Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao,

Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, volume 2024, pp. 2483–2503, 2024.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BH8TYy0r6u.

Thomas Jiralerspong and Trenton Bricken. Cross-architecture model diffing with crosscoders: Unsupervised discovery of differences between llms. *ArXiv preprint*, abs/2602.11729, 2026. URL https://arxiv.org/abs/2602.11729.

Mario Köppen. The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (WSC5)*, volume 1, pp. 4–8, 2000.

Daniel James Korchinski, Dhruva Karkada, Yasaman Bahri, and Matthieu Wyart. On the emergence of linear analogies in word embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VV1HfpDCF1.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32 – 73, 2016. URL https://api.semanticscholar.org/CorpusID:4492210.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *ArXiv preprint*, abs/2309.00267, 2023. URL https://arxiv.org/abs/2309.00267.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Michael Li and Nishant Subramani. Model internal sleuthing: Finding lexical identity and inflectional features in modern language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15825–15864, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.720. URL https://aclanthology.org/2026.acl-long.720/.

Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344, March 2025. ISSN 1099-4300. doi: 10.3390/e27040344. URL http://dx.doi.org/10.3390/e27040344.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, pp. 3982–3992, 2024.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *ArXiv preprint*, abs/2505.24864, 2025. URL https://arxiv.org/abs/2505.24864.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics. *ArXiv preprint*, abs/2403.01509, 2024. URL https://arxiv.org/abs/2403.01509.

Naman Malhotra, Jay Ambadkar, Abhinav Gupta, Kushal Kasivel, Abbas Schwarz, Kamillo Ferry, and Anthea Monod. Tracking representation dynamics in large language models with persistent homology. In *Topology, Algebra, and Geometry in Data Science*, 2026. URL https://openreview.net/forum?id=wFa5yr2f1S.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3195–3204. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00331. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993. URL https://aclanthology.org/H93-1061.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. *ArXiv preprint*, abs/2505.18235, 2025. URL https://arxiv.org/abs/2505.18235.

nostalgebraist. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. LessWrong blog post.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155, 2022. URL https://arxiv.org/abs/2203.02155.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. ISSN 1091-6490. doi: 10.1073/pnas.2015509117. URL http://dx.doi.org/10.1073/pnas.2015509117.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *ArXiv preprint*, abs/2311.03658, 2023. URL https://arxiv.org/abs/2311.03658.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *ArXiv preprint*, abs/2406.01506, 2024. URL https://arxiv.org/abs/2406.01506.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Qihan Ren, Peng Wang, Ruikun Cai, Shuai Shao, Dadi Guo, Yuejin Xie, Yafu Li, Quanshi Zhang, Xia Hu, Jing Shao, and Dongrui Liu. Rethinking generalization in reasoning sft: A conditional analysis on optimization, data, and model capability. *ArXiv preprint*, abs/2604.06628, 2026. URL https://arxiv.org/abs/2604.06628.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in rlvr. *ArXiv preprint*, abs/2506.10947, 2025. URL https://arxiv.org/abs/2506.10947.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv preprint*, abs/2402.03300, 2024. URL https://arxiv.org/abs/2402.03300.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=91H76m9Z94. Survey Certification.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Minghe Shen, Zhuo Zhi, Chonghan Liu, Shuo Xing, Zhengzhong Tu, and Che Liu. Does rlvr extend reasoning boundaries? investigating capability expansion in vision-language models. *ArXiv preprint*, abs/2511.00710, 2025. URL https://arxiv.org/abs/2511.00710.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43): e2200800119, 2022.

Yiming Tang, Arash Lagzian, Srinivas Anumasa, Qiran Zou, Yingtao Zhu, Ye Zhang, Trang Nguyen, Yih-Chung Tham, Ehsan Adeli, Ching-Yu Cheng, et al. Human-like content analysis for generative ai with language-grounded sparse encoders. *npj Artificial Intelligence*, 2025a. URL https://arxiv.org/abs/2508.18236.

Yiming Tang, Harshvardhan Saini, Zhaoqian Yao, Zheng Lin, Yizhen Liao, Jingyi Cui, Yisen Wang, Mengnan Du, and Dianbo Liu. A unified theory of sparse dictionary learning in mechanistic interpretability: Piecewise biconvexity and spurious minima. *ArXiv preprint*, abs/2512.05534, 2025b. URL https://arxiv.org/abs/2512.05534.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *ArXiv preprint*, abs/2506.01939, 2025a. URL https://arxiv.org/abs/2506.01939.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025b. URL https://icml.cc/virtual/2025/poster/46507.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement learning for reasoning in large language models with one training example. *ArXiv preprint*, abs/2504.20571, 2025c. URL https://arxiv.org/abs/2504.20571.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jGbRWwIidy.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *ArXiv preprint*, abs/2502.17420, 2025. URL https://arxiv.org/abs/2502.17420.

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.

Bo Xiong. The lattice representation hypothesis of large language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.01227.

Yang Xu, Yi Wang, Hengguan Huang, and Hao Wang. Tracking the feature dynamics in llm training: A mechanistic study. *ArXiv preprint*, abs/2412.17626, 2024. URL https://arxiv.org/abs/2412.17626.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *ArXiv preprint*, abs/2407.10671, 2024. URL https://arxiv.org/abs/2407.10671.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4OsgYD7em5.

Haiyan Zhao, Zirui He, Yiming Tang, Fan Yang, Ali Payani, Dianbo Liu, and Mengnan Du. Rep2text: Decoding full text from a single llm token representation. *ArXiv preprint*, abs/2511.06571, 2025. URL https://arxiv.org/abs/2511.06571.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning. In *Proceedings of the 39th Conference on Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.01347.

A Related Work

A.1 Lens to Interpret Model Internals

Diverse methods have been proposed to interpret a model's internal representation. The Logit Lens (nostalgebraist, 2020) projects intermediate hidden states through the unembedding to reveal a layerwise prediction trajectory, and the Tuned Lens (Belrose et al., 2023) corrects its basis-mismatch failures with learned per-layer translators. Patchscopes (Ghandeharioun et al., 2024) patches hidden states into inspection prompts to decode them in natural language, and Rep2Text (Zhao et al., 2025) decodes full text from a single token representation. Probing classifiers (Alain & Bengio, 2016; Hewitt & Manning, 2019) test whether predefined concepts are linearly decodable. Sparse autoencoders decompose polysemantic activations into dictionaries of monosemantic features (Huben et al., 2024; Gao et al., 2025; Bussmann et al., 2025), with LanSE grounding features directly in natural language (Tang et al., 2025a) and natural language autoencoders producing unsupervised explanations of activations (Fraser-Taliente et al., 2026). For comparison across models and training, crosscoders diff models through shared dictionaries (Lindsey et al., 2024; Jiralerspong & Bricken, 2026) and SAE Track follows feature dynamics along a training trajectory (Xu et al., 2024). Despite their diversity, all of these lenses map activations onto tokens, language, or predefined and learned features, spaces that are easier to interpret, but none directly reports *how concepts are encoded* in representation space—the spread, anisotropy, and norm structure of the region a concept occupies.

A.2 Neural Geometry

Neural geometry refers to the study of how information is spatially organized in the representation spaces of neural networks, often through overarching hypotheses supported by suggestive yet partial evidence. The linear representation hypothesis holds that high-level concepts are encoded as directions (Park et al., 2023), with roots in the linear analogies of word embeddings (Korchinski et al., 2025); superposition further explains how more features than dimensions can coexist as almost-orthogonal directions (Elhage et al., 2022), and diverse evidence supports this linear view, from refusal directions to linear representations of space and time and emergent world models (Arditi et al., 2024; Gurnee & Tegmark, 2024; Li et al., 2023). Beyond directions, various works uncover richer, nonlinear structures: categorical and hierarchical concepts form polytopes and orthogonal hierarchies (Park et al., 2024), some features are irreducibly multi-dimensional, forming circles for days and months (Engels et al., 2025), SAE feature clouds exhibit multi-scale structure (Li et al., 2025), and manifold steering reveals shared geometry between representation and behavior (Wurgaft et al., 2026). At a global level, neural collapse suggests that class representations might collapse to a simplex structure in the terminal phase of training (Papyan et al., 2020), while the platonic representation hypothesis suggests that representations across models and modalities might converge (Huh et al., 2024). However, we currently lack a method to track how training reshapes neural geometry; such changes remain characterized mostly by hypotheses that are not yet sufficiently supported.

A.3 Understanding How Post-Training (SFT & RL) Works

Supervised fine-tuning (SFT) and reinforcement learning (RL) are the prevailing paradigms of model post-training, widely utilized for human alignment via RLHF and RLAIF (Ouyang et al., 2022; Lee et al., 2023) and for enhancing reasoning via RLVR (Shao et al., 2024; Wang et al., 2025c). Yet how these methods actually work remains poorly understood, and existing works hold seemingly contradictory views. Some argue RLVR merely sharpens the output distribution over solutions the base model can already produce (Yue et al., 2025; Shao et al., 2025) and may even collapse the model's capability boundary (Dong et al., 2025); others find RL genuinely incentivizes correct reasoning (Wen et al., 2026) and that prolonged RL expands reasoning boundaries toward novel capabilities (Liu et al., 2025; Shen et al., 2025). Comparative studies report that SFT tends to memorize training data while RL generalizes to unseen rule and visual variants, in both LLMs and VLMs (Chu et al., 2025). However, others show this narrative is conditional rather than universal, with SFT's generalization depending on optimization choices, data composition, and model capability (Ren et al., 2026). Finer-grained analyses identify key components of RL, such as high-entropy forking tokens (Wang et al., 2025a), and show negative reinforcement mitigates diversity loss (Zhu et al., 2025). Crucially, almost all of these works provide perspectives at the behavior level, measuring shifts in the output distributions, and do not answer what RL and SFT do to the representation space itself.

B DISCUSSION

Tracking representation geometry. Mechanistic interpretability has largely focused on characterizing representations at a fixed model checkpoint, through directions, features, probes, or geometric structure (Park et al., 2023; Engels et al., 2025; Lindsey et al., 2024; Jiralerspong & Bricken, 2026; Duan et al., 2026). Comparatively less attention has been paid to how the geometry of those representations evolves throughout training, although recent work has begun to study feature and circuit dynamics across checkpoints (Xu et al., 2024; Wang et al., 2025b). This question is increasingly important as modern models undergo multiple stages of pretraining and post-training whose effects on internal representations remain only partially understood (Chu et al., 2025; Yue et al., 2025; Wen et al., 2026). We therefore view tracking representation geometry as a complementary problem to interpreting a static representation: rather than asking only what a model represents, we also ask where and how that representation changes as the model is trained. Capsule Lens provides a simple way to make such changes measurable by matching the same interpretable geometric parameterization across model states.

Coverage and disentanglement as recall and precision. The two metrics used in this work admit a natural interpretation analogous to recall and precision. Held-out coverage measures the fraction of unseen positive samples captured by a fitted capsule and can therefore be viewed as the *recall* of the capsule in locating a concept's geometry. Conversely, the disentanglement score measures the probability that a representation admitted by a capsule belongs to its target concept and can be viewed as its *precision*. This distinction is important because prior work suggests that concept representations need not be perfectly isolated or one-dimensional, but may instead occupy overlapping, multidimensional, or manifold-like regions (Sorscher et al., 2022; Engels et al., 2025; Modell et al., 2025; Bhalla et al., 2026). Since a capsule is an intentionally simple geometric family fitted in closed form, high held-out coverage indicates that the fitted region generalizes to unseen instances of the concept, whereas high disentanglement indicates that this simple region is relatively specific to that concept. We therefore view disentanglement not only as a property of the fitting procedure, but also as a characteristic of the underlying concept geometry: concepts whose representations are compactly approximated by a capsule should naturally achieve higher precision than concepts whose geometry is strongly intertwined with other concepts.

Why capsules? Our choice of capsules is motivated primarily by simplicity and interpretability rather than by the claim that concept geometry must take this form. Prior work has described concept representations using directions (Park et al., 2023; Arditi et al., 2024), higher-dimensional manifolds (Sorscher et al., 2022; Engels et al., 2025; Modell et al., 2025), categorical and hierarchical geometric structures (Park et al., 2024), and cone-like regions (Wollschläger et al., 2025). Capsules draw inspiration from these geometric views while retaining only a small number of directly interpretable parameters: the axis describes orientation, the axis span describes angular spread, and the norm bounds describe radial extent. These parameters can be computed in closed form and compared directly across layers, models, and training checkpoints. More expressive geometric descriptions, including ellipsoidal or topological characterizations (Sorscher et al., 2022; Malhotra et al., 2026), may capture richer structure but are less immediately reducible to a small set of quantities whose changes can be interpreted across model states. Capsule Lens therefore deliberately trades geometric flexibility for a compact and trackable description of representation. We also compare capsules with alternative geometric forms in terms of held-out coverage and disentanglement, providing empirical guidance for choosing different geometric approximations when appropriate.

Sensitivity to boundary samples. The closed-form capsule matching procedure defines its angular and norm boundaries using extrema of the fitting samples, making the resulting axis span and norm bounds potentially sensitive to atypical boundary samples. We address this issue systematically in Appendix D.2 by replacing the minimum and maximum operations with percentile-based boundaries and evaluating the resulting trade-off between held-out coverage and disentanglement. These experiments show that the qualitative conclusions remain stable across boundary choices, while tighter percentile boundaries improve disentanglement at the cost of excluding more unseen positive samples. The span and norm curves provide an additional way to diagnose such sensitivity by revealing how representations populate the interior of the fitted capsule, including whether the fitted boundary is determined by only a small number of extreme samples. In practice, we there-

fore recommend two complementary uses of Capsule Lens. When the internal geometric structure is of interest, the original min/max construction can be used together with the coverage curves, which preserve the full observed extent while exposing how densely different parts of the region are occupied. When only the capsule parameters themselves are required, percentile-based boundaries provide a natural robust alternative whose strictness can be chosen according to the desired coverage–disentanglement trade-off. Detailed comparisons across percentile choices are provided in Appendix D.2.

From case studies to systematic hypotheses. The three dynamic case studies reveal several geometric patterns that merit substantially more systematic investigation. CLIP pretraining produces broad restructuring across the network, PPO post-training on visual question answering produces much smaller and strongly depth-localized changes, and RLVR on mathematical reasoning reveals sparse concept- and layer-specific concentration. These observations connect to a growing literature showing that pretraining, fine-tuning, and reinforcement learning can affect model representations in qualitatively different ways (Chu et al., 2025; Xu et al., 2024; Liu et al., 2025; Yue et al., 2025; Wen et al., 2026; Wang et al., 2025a). They also raise broader questions about which geometric changes are associated with behavioral improvements, whether different training objectives produce characteristic geometric signatures, and whether these patterns generalize across models, datasets, optimization settings, and random seeds. Establishing such general laws would require dedicated experimental studies for each training regime and is beyond the scope of the present work. Our goal here is instead to demonstrate that Capsule Lens provides a practical instrument for posing and studying these questions. We therefore treat the findings of our case studies as empirical hypotheses suggested by the framework and leave their systematic validation to future work.

C Future Work

We highlight four promising directions for future work:

- **Systematically tracking specific training pipelines.** Capsule Lens can be used to study in detail how particular training procedures reshape concept geometry. Extending the case studies in this work into dedicated analyses across objectives, datasets, checkpoints, and model families can provide key insights to better understand different training pipelines.
- **Characterizing geometric differences across model layers.** Capsule Lens can be used to systematically compare concept geometry across depth, helping reveal how concepts are formed, transformed, concentrated, or disentangled between different layers of a model.
- **Studying hierarchical concepts through capsule intersections.** Capsule Lens may offer a geometric view of concept hierarchy by analyzing how capsules corresponding to related concepts intersect, overlap, or contain one another. This could provide a direct way to investigate how hierarchical semantic structure is organized in representation space.
- **Steering representations with capsules.** Beyond analysis, the geometric structure identified by Capsule Lens may provide a basis for representation steering. Interventions along a capsule's axis, within its angular span, or toward particular regions of its norm band could offer a geometrically grounded way to manipulate concept-related representations and study their causal effects on model behavior.

D SYSTEMATICALLY VALIDATING CAPSULE LENS

We provide systematic experiments further validating the reliability and design choices of Capsule Lens. Section D.1 examines the effect of the number of samples used for capsule matching and the stability of the estimated geometry under sampling. Section D.2 evaluates percentile-based capsule boundaries in place of the minimum and maximum fitting samples, examining the resulting trade-off between held-out coverage and disentanglement. Section D.3 further compares capsules with alternative geometric approximations across different boundary choices.

D.1 EFFECT OF THE NUMBER OF SAMPLES PER CONCEPT

We examine how the number of samples used for capsule matching affects the estimated geometry. We vary m from 5 to 100 while evaluating each fitted capsule on the same set of 100 held-out samples. The experiment covers 796 visual concepts across eight model components, using three representative layers from each component and three independent draws per concept.

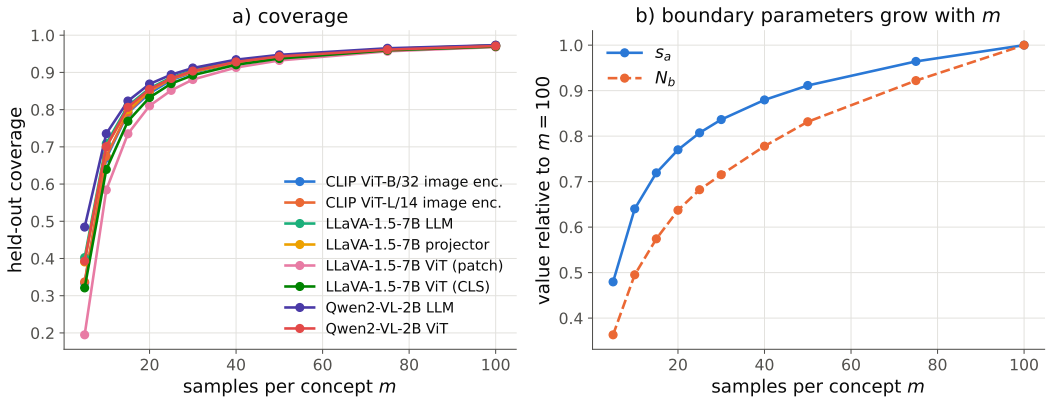


Figure 7: **Effect of the number of samples m on fitted capsules.** Results are averaged over 796 visual concepts, eight model components, three representative layers per component, and three independent draws. **a)** Held-out coverage increases with m and gradually saturates as more samples are used for capsule matching. **b)** Axis span s_a and norm band N_b, normalized by their values at $m = 100$, increase with m as the fitted capsule progressively covers the concept geometry.

Held-out coverage increases rapidly with m, reaching 0.878 at $m = 25$, 0.941 at $m = 50$, and 0.971 at $m = 100$ (Figure 7). The fitted axis span and norm band also increase with m, as larger samples progressively capture more of the concept's angular and norm extent. In particular, the $m = 50$ fitting size used in our main experiments already achieves a mean held-out coverage of 0.941, while increasing the sample size further provides diminishing gains.

We further evaluate the sampling stability of the fitted axis by independently matching two capsules to disjoint sets of m samples from the same concept and measuring the cosine distance Δa between their axes. The estimated axes become increasingly consistent as m grows, with mean Δa decreasing from 0.1150 at $m = 5$ to 0.0141 at $m = 50$ and 0.0071 at $m = 100$. Table 4 further shows that this trend is consistent across model components. These values provide a reference scale for interpreting the axis drifts measured across training checkpoints.

D.2 ROBUSTNESS TO CAPSULE BOUNDARY ESTIMATION

Capsule matching defines its boundary using the extrema of the fitting samples: $s_a = \max_i d_{\cos}(x_i, a)$, $n_{\text{low}} = \min_i \|x_i\|$, and $n_{\text{high}} = \max_i \|x_i\|$. To examine whether our results depend on this choice, we replace these extrema with percentile-based boundaries parameterized by q. We set s_a to the q-th percentile of the angular distances and $(n_{\text{low}}, n_{\text{high}})$ to the central $q\%$ interval of the sample norms, while keeping the fitted axis unchanged. We evaluate $q \in \{100, 95, 90, 85, 80\}$, where $q = 100$ recovers the original capsule. We perform this analysis across different types of concepts and evaluate each fitted capsule using held-out coverage and disentanglement Dis.

Table 4: **Held-out coverage, axis span, and axis dispersion across model components.** Results are computed over 796 visual concepts and three representative layers per component. Axis dispersion Δa is the cosine distance between axes independently fitted from two disjoint sample sets of the same concept.

		Coverage			s_a		Δa	
Model	Component	m=25	m=50	m=100	m=25	m=100	m=25	m=100
CLIP ViT-B/32	image enc.	0.879	0.942	0.972	0.146	0.175	0.0107	0.0027
CLIP ViT-L/14	image enc.	0.882	0.943	0.972	0.214	0.265	0.0142	0.0036
Qwen2-VL-2B	ViT	0.884	0.943	0.971	0.412	0.530	0.0277	0.0072
Qwen2-VL-2B	LLM	0.894	0.947	0.973	0.398	0.478	0.0176	0.0044
LLaVA-1.5-7B	ViT (patch)	0.852	0.932	0.968	0.646	0.745	0.0739	0.0198
LLaVA-1.5-7B	ViT (CLS)	0.870	0.937	0.969	0.273	0.323	0.0252	0.0065
LLaVA-1.5-7B	projector	0.884	0.943	0.973	0.507	0.640	0.0271	0.0068
LLaVA-1.5-7B	LLM	0.887	0.944	0.972	0.482	0.665	0.0231	0.0058

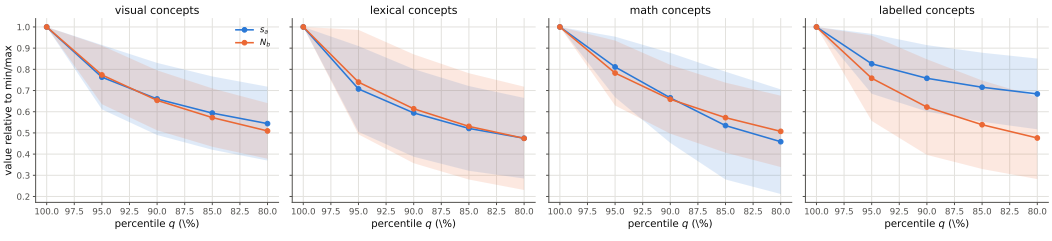


Figure 8: **Capsule parameters under percentile-based boundaries.** Axis span s_a and norm band N_b are normalized by their values at $q = 100$ and averaged across concepts. Shaded regions indicate ± 1 standard deviation. Lower percentiles produce progressively tighter capsule boundaries across all four concept types.

As shown in Figure 8, both the axis span and norm band decrease consistently as the percentile is lowered. The largest contraction occurs between $q = 100$ and $q = 95$: at $q = 95$, the mean axis span is 0.708–0.826 of its original value across concept types, while the mean norm band is 0.740–0.783 of its original value. Thus, the extreme samples have a substantial effect on the fitted boundary, but the same qualitative contraction is observed across modalities and concept constructions.

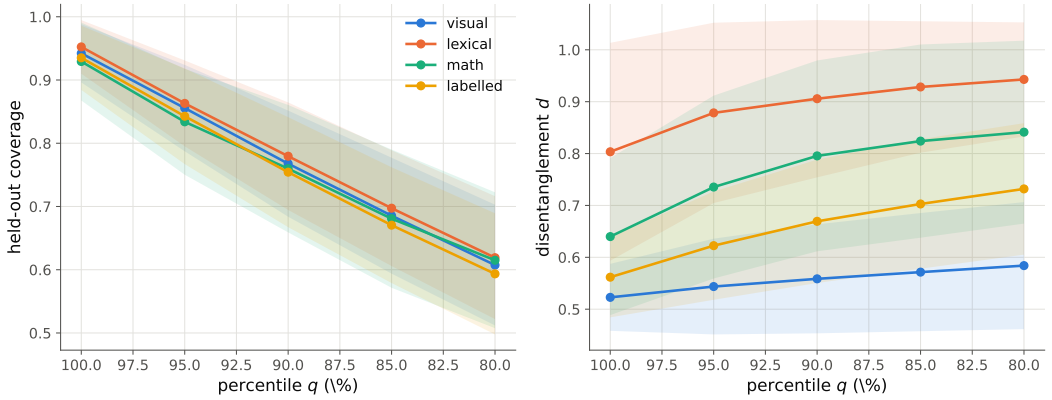


Figure 9: **Held-out coverage and disentanglement under percentile-based boundaries.** Held-out coverage (left) and disentanglement Dis (right) as the percentile q decreases from the original min/max boundary ($q = 100$) to increasingly tighter capsules. Results are averaged across concepts, with shaded regions indicating ± 1 standard deviation.

Tighter boundaries induce a consistent trade-off between coverage and disentanglement (Figure 9). Reducing q from 100 to 80 decreases held-out coverage by approximately 0.32–0.34 across all four

concept types, while increasing disentanglement by 0.061 for visual, 0.140 for lexical, 0.201 for math, and 0.170 for labelled concepts. These results show that percentile-based boundaries can substantially increase the selectivity of fitted capsules, but only at the cost of excluding a considerable fraction of unseen positive samples. We therefore retain the min/max boundary in Capsule Lens, prioritizing generalization to unseen samples while reporting disentanglement separately to characterize overlap with other concepts.

D.3 Alternative Geometric Approximations under Different Boundaries

We further examine whether the comparison with alternative geometric approximations in Section 3.3 persists under percentile-based boundaries. A *cone* retains the capsule axis a and angular span s_a without constraining the norm, while a *band* retains only the norm interval $[n_{\text{low}}, n_{\text{high}}]$. A *ball* is centered at the mean of the fitting samples with radius $r = \max_i \|x_i - c\|$, and a *box* uses the coordinate-wise minima and maxima of the fitting samples. All shapes are fitted to the same 50 samples and evaluated using the same concepts, layers, held-out samples, and negatives.

We apply the percentile-based boundary construction of Section D.2 to each geometric approximation. Table 5 reports held-out coverage and disentanglement across $q \in \{100, 95, 90, 85, 80\}$. Tightening the boundaries produces the expected coverage–disentanglement trade-off for the capsule, cone, and ball. Across all five percentiles, the capsule consistently achieves higher disentanglement than the cone and ball, while the cone and ball retain higher coverage at the same q. The band remains substantially less disentangled across all boundary choices, with Dis increasing only from 0.514 to 0.532, further indicating that norm information alone provides limited separation. The box continues to exhibit extremely low held-out coverage across all settings. Overall, the results reveal a consistent pattern across boundary choices: directional information accounts for most of the observed concept separation, while incorporating the norm constraint yields a modest additional increase in disentanglement. The capsule therefore provides a compact parameterization that jointly captures both aspects while maintaining high held-out coverage under its original boundary.

Table 5: **Alternative geometric approximations under percentile-based boundaries.** Held-out coverage (Cov.) and disentanglement (Dis.), averaged across concept types, model components, and representative layers. Lower q corresponds to a tighter boundary; $q = 100$ recovers the min/max construction.

	$q = 100$		$q = 95$		$q = 90$		$q = 85$		$q = 80$	
Shape	Cov.	Dis.	Cov.	Dis.	Cov.	Dis.	Cov.	Dis.	Cov.	Dis.
Capsule	0.943	**0.583**	0.855	**0.620**	0.769	**0.641**	0.686	**0.658**	0.609	**0.673**
Cone	0.975	0.578	0.918	0.614	0.862	0.635	0.805	0.651	0.750	0.665
Ball	0.977	0.568	0.922	0.601	0.865	0.620	0.811	0.635	0.757	0.649
Band	0.962	0.514	0.916	0.521	0.867	0.526	0.818	0.529	0.769	0.532
Box	0.012	0.653	0.003	0.456	0.003	0.890	0.002	1.000	0.002	1.000

E CASE STUDY I: TRACKING CONCEPT GEOMETRY IN CLIP PRETRAINING

This appendix extends the main-text discussion of Section 5.1 and Figure 4, giving additional views on the same training run and checkpoints — the within-concept distribution behind the block-11 span contraction, and the concept-level analysis of the drift signal.

E.1 AXIS-SPAN CONTRACTION IS POPULATION-WIDE

Figure 10 shows that the block-11 span contraction reported in the main text holds across the concept population, rather than being driven by a subset of concepts.

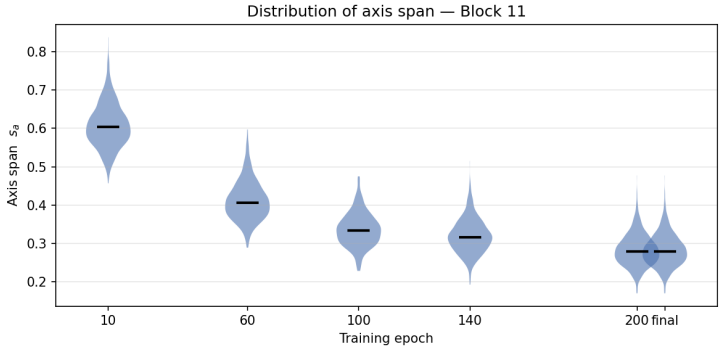


Figure 10: **Block-11 span contraction is visible across the concept distribution.** Distribution of axis-span s_a across concepts at block 11, by training epoch. The distribution shifts downward (mean $\approx 0.60 \rightarrow 0.28$) and narrows over training, without a separate non-contracting subpopulation appearing at any checkpoint.

E.2 CONCEPT-LEVEL STRUCTURE

Figures 11–12 break down the block-11 rotation reported in the main text by concept, illustrating that Capsule Lens can rank individual concepts by drift magnitude within a single probed block.

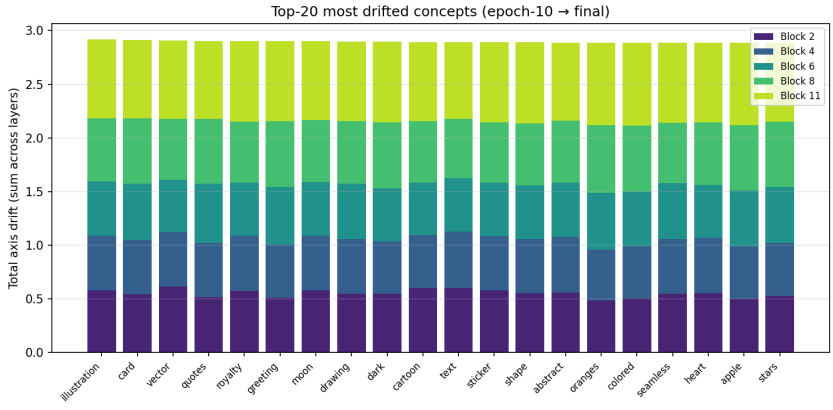


Figure 11: **Top-20 most drifted concepts, block 11.** Total axis drift (epoch-10 $\rightarrow$ final) summed across the five probed blocks (stacked bars, one color per block), for the 20 concepts with the largest totals. Totals span a narrow range ($\approx$ 2.88–2.92).

E.3 AXIS DRIFT VS. SPAN CHANGE

Figure 13 relates the block-11 span contraction to the per-concept drift of Section E.2, jointly plotting both quantities for individual concepts.

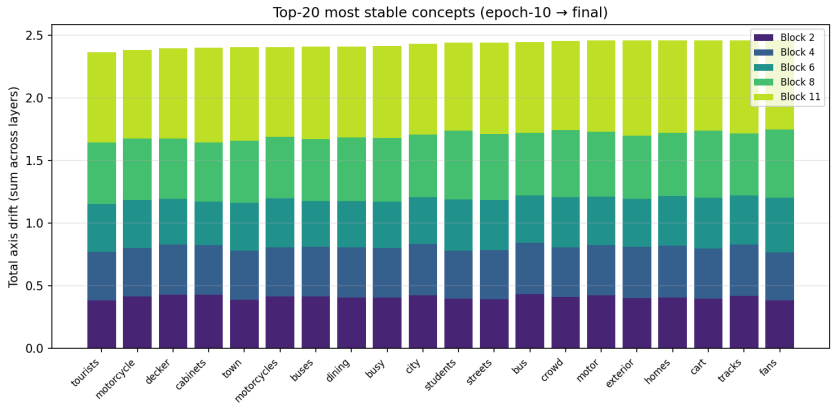


Figure 12: **Top-20 most stable concepts, block 11.** Total axis drift (epoch-10 $\rightarrow$ final) summed across the five probed blocks, for the 20 concepts with the smallest totals. Totals ($\approx$ 2.36–2.46) are only $\approx 20\%$ below the most-drifted group in Figure 11, indicating that the gap between the two extremes is present but modest relative to the overall drift scale.

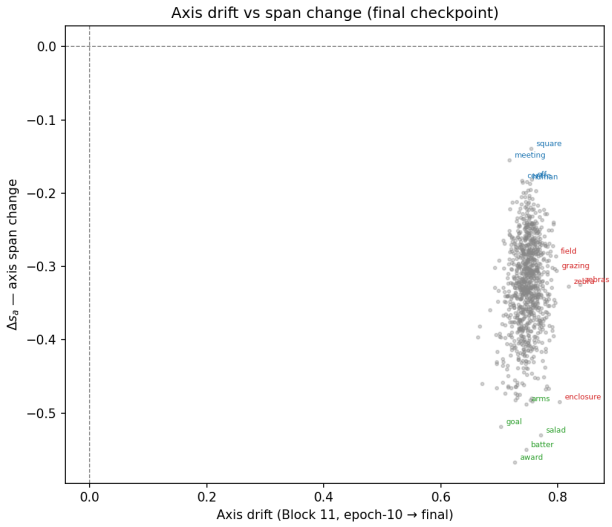


Figure 13: **Axis drift and span change at block 11, final checkpoint.** Axis drift vs. span change Δs_a for individual concepts, with five example concepts labeled at each of three regions of the plot: lowest span change in magnitude (blue: *square*, *meeting*, *coffee*, *human*), highest axis drift (red: *field*, *grazing*, *zebra*, *zebras*, *enclosure*), and largest span change in magnitude (green: *arms*, *goal*, *salad*, *batter*, *award*). All plotted concepts fall in a narrow drift band (≈ 0.65–0.87) and have $\Delta s_a < 0$ ($\Delta s_a \in [-0.57, -0.14]$).

F CASE STUDY II: PPO FOR VLM ON VISUAL QUESTION ANSWERING

This appendix extends the main-text discussion of PPO post-training on Qwen2-VL, giving additional views on the same training run and checkpoints — the concept-level breakdown of the layer-27 drift signal, and the joint relationship between axis drift and span change.

F.1 CONCEPT-LEVEL HETEROGENEITY

Figures 15–17 rank individual concepts by total axis drift at layer 27, the layer where drift concentrates.

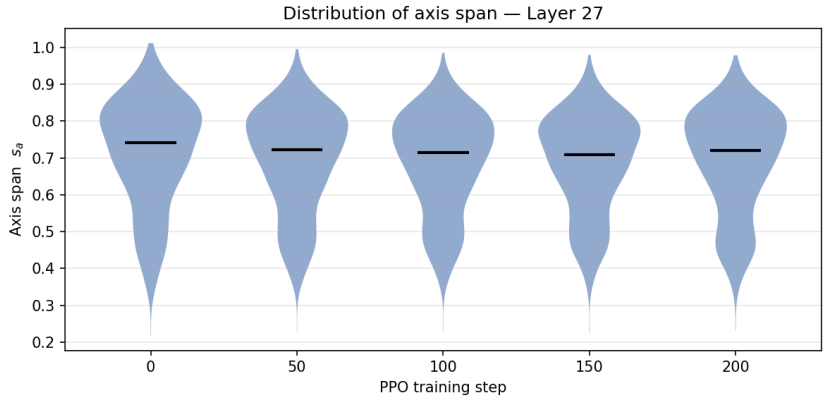


Figure 14: **Axis-span distribution at layer 27 across checkpoints.** Distribution of axis-span s_a across concepts at layer 27, by PPO training step. The distribution shape is similar across all five checkpoints; the mean varies between ≈ 0.71 and ≈ 0.74.

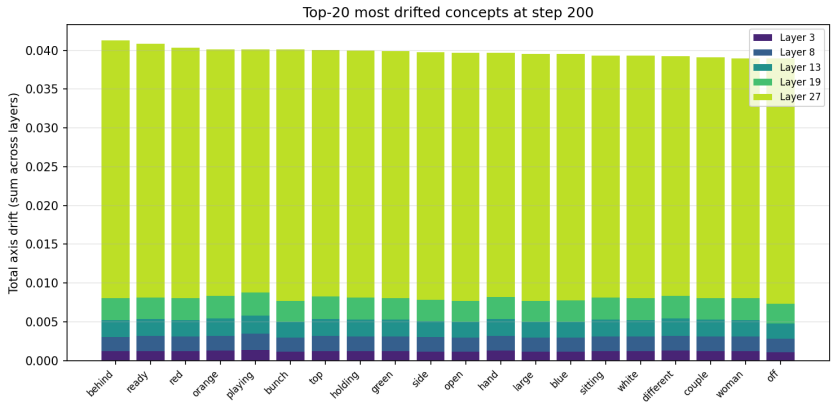


Figure 15: **Top-20 most drifted concepts, layer 27.** Total axis drift at step 200, summed across the five probed layers (stacked bars, one color per layer), for the 20 concepts with the largest totals. Layer 27 (top segment) accounts for most of each total.

F.2 AXIS DRIFT VS. SPAN CHANGE

Figure 16 jointly plots axis drift and span change Δs_a for individual concepts at layer 27, the final checkpoint.

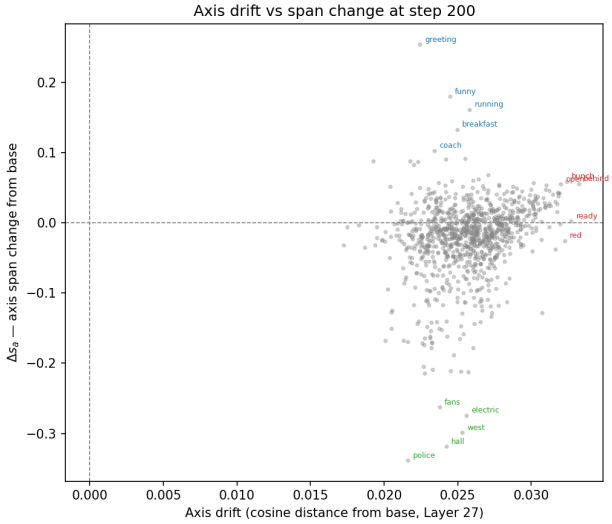


Figure 16: **Axis drift and span change at layer 27, final checkpoint.** Axis drift vs. span change Δs_a (step 200 minus base) for individual concepts, with five example concepts labeled at each of three regions of the plot: largest positive span change (blue: *greeting*, *funny*, *running*, *breakfast*, *coach*), lowest span change in magnitude (red: *bunch*, *open*, *behind*, *ready*, *red*), and largest negative span change (green: *fans*, *electric*, *west*, *hall*, *police*). Axis drift for all plotted concepts falls between 0.018 and 0.033.

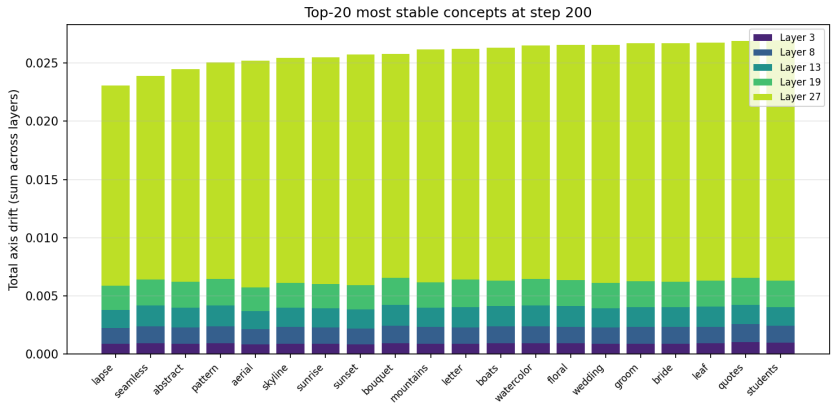


Figure 17: **Top-20 most stable concepts, layer 27.** Total axis drift at step 200, summed across the five probed layers, for the 20 concepts with the smallest totals. The y-axis maximum (≈ 0.027) is close to Figure 15's (≈ 0.041), indicating the gap between most- and least-drifted concepts is present but modest.

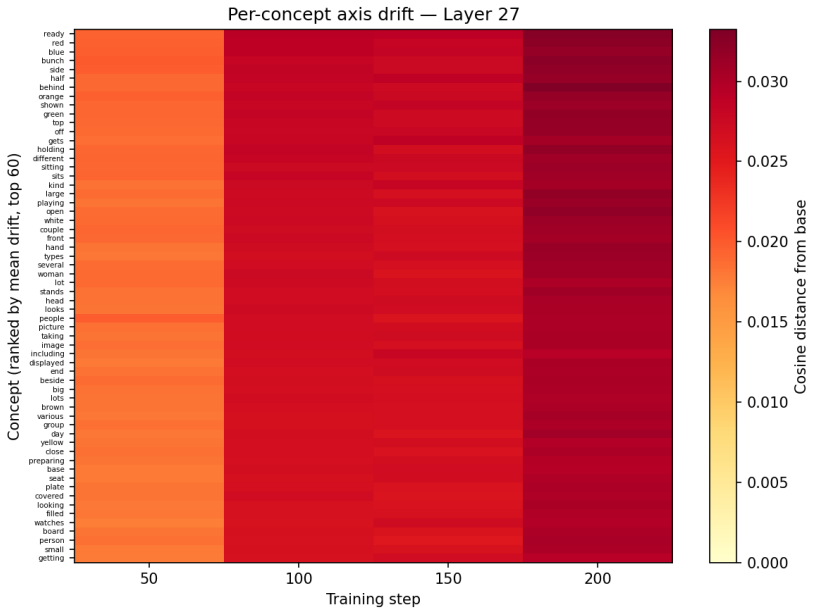


Figure 18: **Per-concept axis drift at layer 27 across training.** Rows are the top 60 concepts by mean drift, columns are PPO training steps. Color encodes cosine distance from the base axis.

G Case Study III: RLVR on Mathematical Reasoning

This appendix provides additional analyses for the RLVR experiment in Section 5.3. We examine both the population-level evolution of concept geometry and the individual trajectories of all 18 mathematical concepts under the same GRPO training run. Capsules are fitted using 50 samples per concept and evaluated on a disjoint set of 50 samples; held-out coverage averages 0.942 across training, ranging from 0.80 to 1.00.

G.1 Population-Averaged Geometry

We first average each capsule parameter across the 18 mathematical concepts at every probed layer. Figure 19 shows that population-level geometry remains highly stable under RLVR. Mean axis drift stays below 7×10^{-6} at every layer, and axis spans remain nearly unchanged at layers 3–20; only layer 27 exhibits a small non-monotonic change followed by a modest contraction. Mean norm-band midpoint and width are similarly stable throughout training. These averages therefore indicate that RLVR induces little global restructuring of concept geometry.

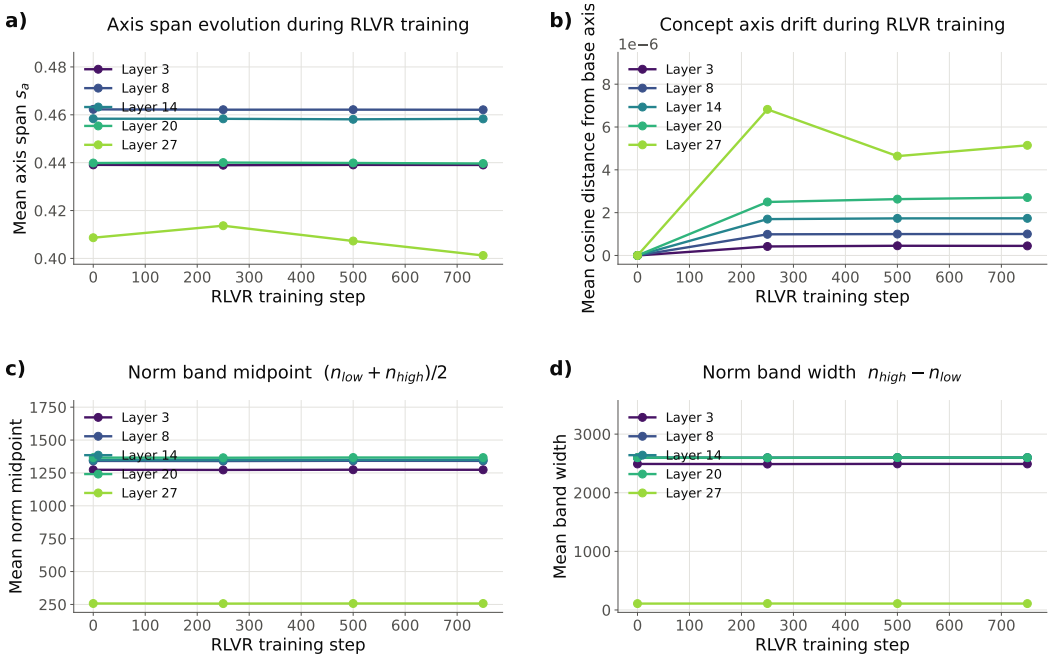


Figure 19: **Population-averaged concept geometry during RLVR. (a)** Mean axis span across the 18 mathematical concepts remains nearly constant at layers 3–20, while layer 27 shows a small late contraction. **(b)** Mean axis drift from the base model remains below 7×10^{-6} at every layer. **(c,d)** Mean norm-band midpoint and width remain largely unchanged throughout training.

The population average can nevertheless conceal changes concentrated in a small number of concepts. We therefore inspect the trajectory of each mathematical concept separately.

G.2 Per-Concept Trajectories

Per-concept trajectories reveal a strongly heterogeneous response to RLVR. At layer 27, only 7 of the 18 concepts change their axis span by more than 0.005, while `unit_time_minute` exhibits by far the largest contraction and is analyzed in Figure 6 of the main text. The following figures report axis drift, axis span, norm-band midpoint, and norm-band width for the remaining concepts at all five probed layers.

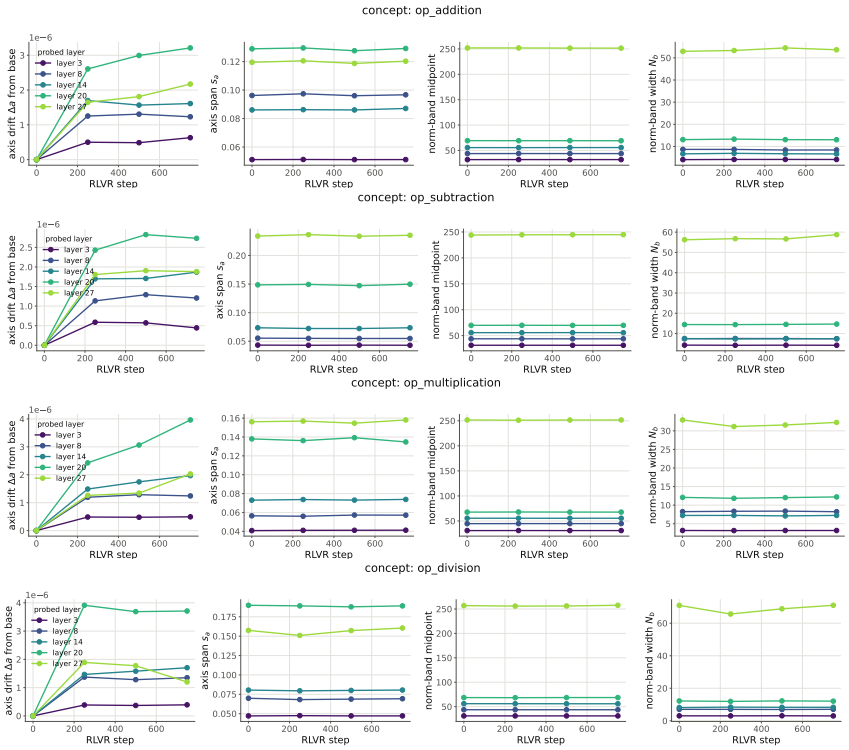


Figure 20: **RLVR trajectories of arithmetic-operation concepts (I).** The concepts `op_addition`, `op_subtraction`, `op_multiplication`, and `op_division` remain largely stable across training, with only minor changes in axis span and norm geometry.

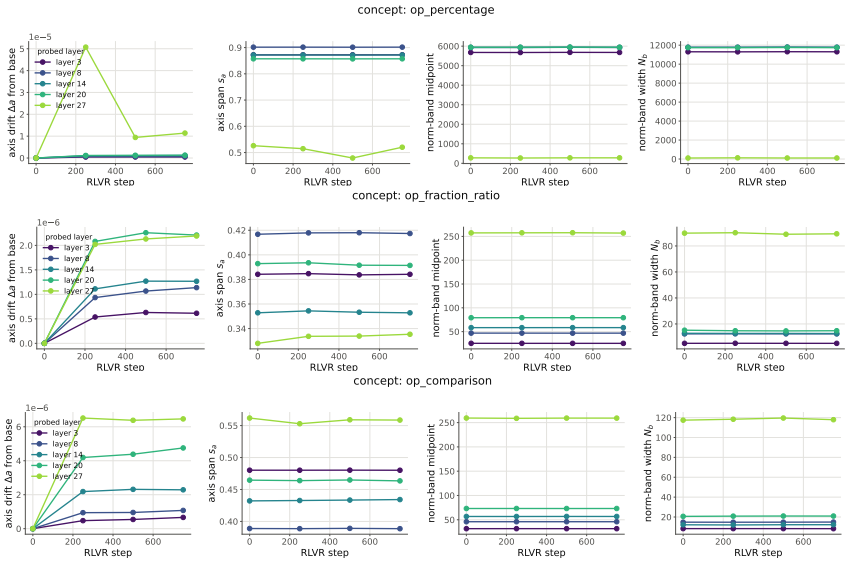


Figure 21: **RLVR trajectories of arithmetic-operation concepts (II).** The concepts `op_percentage`, `op_fraction_ratio`, and `op_comparison` also exhibit limited geometric change, although `op_percentage` shows somewhat larger variation than most other operation concepts. Overall, arithmetic-operation concepts remain geometrically stable under RLVR.

Taken together, these trajectories explain why the population averages remain nearly unchanged despite several visible outliers: RLVR does not broadly restructure the mathematical concept space, but selectively modifies the geometry of a small number of concepts at particular layers. This provides additional evidence for the concept- and layer-specific concentration effect highlighted in Section 5.3.

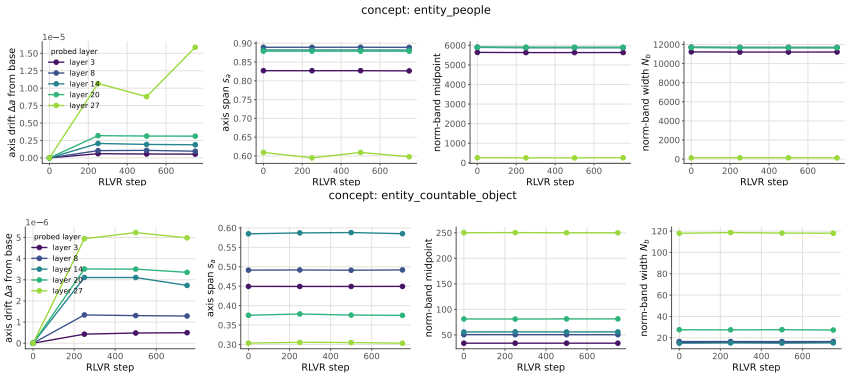


Figure 22: **RLVR trajectories of entity concepts.** `entity_people` exhibits one of the clearest geometric changes outside `unit_time_minute`, including a layer-27 span contraction and substantial norm-bound changes at intermediate layers. In contrast, `entity_countable_object` remains comparatively stable, further illustrating the concept-specific nature of RLVR-induced geometric change.

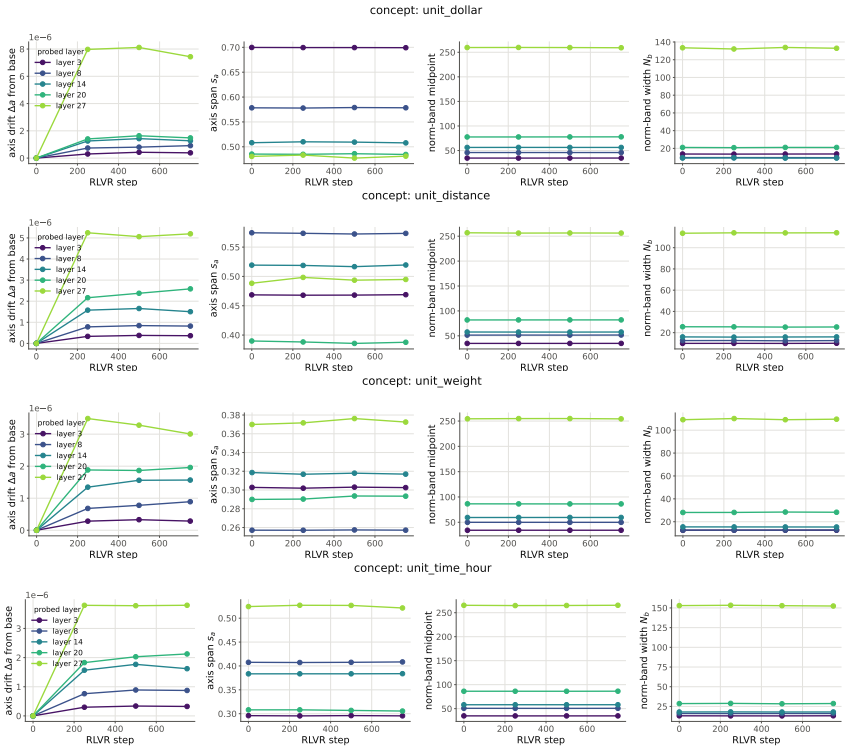


Figure 23: **RLVR trajectories of unit concepts (I).** The concepts `unit_dollar`, `unit_distance`, `unit_weight`, and `unit_time_hour` remain largely stable throughout training, with only small changes in axis span and norm geometry.

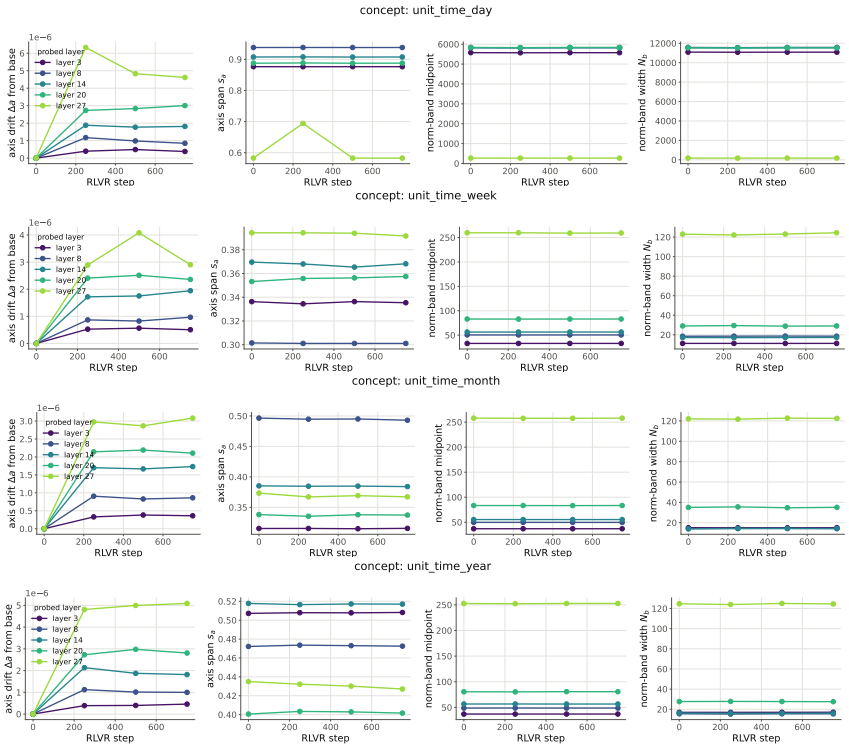


Figure 24: **RLVR trajectories of unit concepts (II).** The concepts `unit_time_day`, `unit_time_week`, `unit_time_month`, and `unit_time_year` exhibit heterogeneous but generally limited changes. Among them, `unit_time_month` and `unit_time_year` show the largest remaining layer-27 span contractions, but both are substantially smaller than the contraction observed for `unit_time_minute` in the main text. This further indicates that RLVR-induced concentration is concept-specific rather than shared uniformly across the time-unit family.

H GEOMETRY-GUIDED REPRESENTATION STEERING WITH CAPSULE LENS

Beyond locating and tracking concept geometry, we test whether the geometry identified by Capsule Lens can also provide useful directions for representation steering. For a concept whose span curve exhibits multimodal structure, we apply spherical k-means to its unit-normalized representations to obtain mode centroids. Given source and target centroids μ_i and μ_j, we define the unit steering direction $d_{i\to j} = (\mu_j - \mu_i)/\|\mu_j - \mu_i\|$. We use the fitted norm shell to calibrate the intervention magnitude, setting $m = (n_{\text{low}} + n_{\text{high}})/2$ and $\lambda = km$ with $k \in [0.6, 0.9]$, and intervene as $h' = h + \lambda d_{i\to j}$. When the two centroids correspond to different modes within one concept, we call this *intra-concept steering*; when they correspond to different concepts, we call it *inter-concept steering*.

Intra-concept steering. Table 6 shows representative examples for polysemous textual concepts. Steering toward opposite modes can selectively elicit different meanings of the same concept, suggesting that the modes identified within a capsule correspond to behaviorally meaningful semantic structure.

Table 6: **Intra-concept steering between semantic modes.** Representative generations from Qwen2.5-1.5B after steering toward different modes within the same concept geometry.

Concept	Target mode	Steered generation
church	institution	"a community of all the members of Christ... authority in the Church"
church	building	"a big, old building... red door with a clock on it"
plant	factory	"a facility for processing... waste gas emissions for steel production"
plant	organism	"long, bushy stems... grow up to 10 inches tall in their garden"
state	nation	"located in the United States... population"
state	condition	"the state of a system... an evolving and dynamic one"
cell	programming	"a variable in the language of programming"

Inter-concept steering. We further construct directions between centroids belonging to different concepts. Table 7 shows that the same geometric construction can shift generations toward distinct textual and visual concepts.

Table 7: **Inter-concept steering across textual and visual representations.** Each example is shown with its base generation followed by the generation obtained after steering toward the target concept.

Modality	Direction	Generation
Text	church → home	**Base:** "a place of worship... pray, sing hymns" **Steered:** "a place where people can relax and enjoy comfort..."
Text	cell → plant	**Base:** "a small unit of life... DNA and proteins" **Steered:** "it has its own life... doesn't need anything else to grow"
Vision	dog → cat	**Base:** "two dogs resting on a cushion in a car" **Steered:** "two cats resting on a fluffy surface..."
Vision	dog → cat	**Base:** "two dogs in the back seat of a car..." **Steered:** "two cats, one orange and one black-and-white..."
Vision	animal → vehicle	**Base:** "two dogs in the back seat of a car..." **Steered:** "two vehicles parked on the side of the road..."
Vision	animal → vehicle	**Base:** "two dogs on a cushion in a car" **Steered:** "a parked vehicle... a large black van with a trailer"

Together, these qualitative results suggest that the geometry recovered by Capsule Lens can provide actionable directions for representation intervention: internal modes support steering between different meanings of the same concept, while relative concept locations support steering across concepts. We leave systematic evaluation of this capability to future work.