From Directions to Cones: Multidimensional Representations of Propositional Facts in LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) exhibit strong conversational abilities but often generate falsehoods. Prior work suggests that the truthfulness of simple propositions can be represented as a single linear direction in a model's internal activations, but this may not fully capture its underlying geometry. In this work, we extend the concept cone framework, recently introduced for modeling refusal, to the domain of truth. We identify multi-dimensional cones that causally mediate truth-related behavior across multiple LLM families. Our results are supported by three lines of evidence: (i) causal interventions reliably flip model responses to factual statements; (ii) learned cones generalize across model architectures; and (iii) cone-based interventions preserve unrelated model behavior. These findings reveal the richer, multidirectional structure governing simple true/false propositions in LLMs and highlight concept cones as a promising tool for probing abstract behaviors.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks, including machine translation, question answering, summarization, code generation, and dialogue systems (Brown et al., 2020; Raffel et al., 2020; Zhang et al., 2020; OpenAI, 2023). Despite their successes, these models remain largely "black boxes" with billions of parameters interacting in complex ways that evade straightforward analysis (Casper et al., 2024). This presents challenges for ensuring alignment with human values and addressing vulnerabilities to adversarial attacks (Hendrycks et al., 2023; Ngo et al., 2022; Hendrycks and Mazeika, 2022). As these models are widely deployed in real-world applications, concerns about reliability and safety have driven a growing interest in model transparency (OpenAI, 2022; Olah et al., 2020; Nanda et al., 2023). Specifically, identifying how and why specific linguistic or behavioral features are encoded within these models is one of the central questions for mechanistic interpretability research (Bereska and Gavves, 2024).

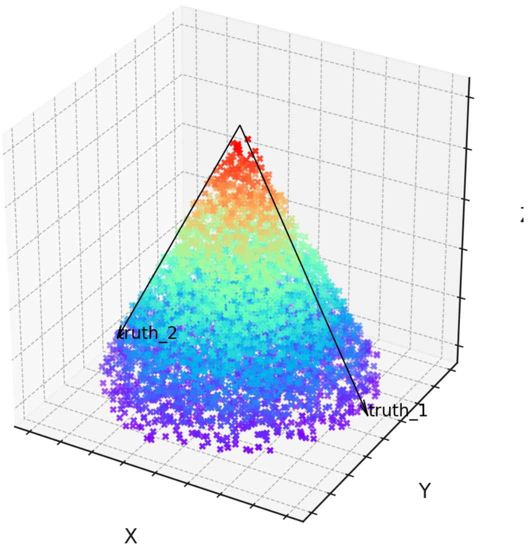

Figure 1: Theoretical visualization of a 2D concept cone. All directions in cone should causally mediate truthful behavior. Given a true propositional input (e.g., "Paris is the capital of France"), ablating along any basis vector of this cone disrupts the model's ability to generate a truthful response. For example, the model will respond with "No, Lyon is the capital" instead.

To analyze the internal representations of LLMs, causal methods such as activation steering and directional ablation (Turner et al., 2024) are used to verify whether modifying specific internal directions leads to corresponding changes in model behavior (Panickssery et al., 2023; Chen et al., 2024). Together, probing and causal interventions have provided insight into how abstract features manifest in model representations.

Previous interpretability studies (Park et al., 2024b,a) have revealed that many high-level features in LLMs correspond to linear directions in the representation space, such as time (Gurnee and Tegmark, 2024), truth (Marks and Tegmark, 2024; Azaria and Mitchell, 2023), space (Gurnee and Tegmark, 2024), political perspective (Kim et al., 2025), and instruction-following (Heo et al., 2025). Other features such as sentiment (Tigges et al., 2023) and refusal (Arditi et al., 2024) have also been shown to exist linearly, although through a different interpretability method known as difference-in-means (DIM). However, the underlying representations may be non-linear, and linear methods may only provide an approximation of some more complex structures (Bürger et al., 2024; Hildebrandt et al., 2025; Engels et al., 2024). Recent work has developed more sophisticated non-linear frameworks and found multiple latent dimensions that capture the fundamental high-level concepts, for refusal in particular (Hildebrandt et al., 2025; Wollschläger et al., 2025).

With sparse autoencoders and concept cones, researchers have characterized multi-dimensional representations of abstract features (Cunningham et al., 2023; Liu et al., 2023; Sharkey et al., 2023). Concept cones, in particular, are a gradient-based search algorithm that, given an initial set of candidate vectors, learn a specific behavior. Each vector is validated to causally influence the target concept through steering or ablation. This nonlinear method extends the interpretability toolkit beyond linear assumptions by enabling both analysis and controlled intervention (Liu et al., 2023; Wollschläger et al., 2025).

In this paper, we extend the concept cone framework to the domain of propositional fact, a subcategory of truthfulness, exploring how this property is internally represented by large language models. Specifically, by applying this framework, we identified a possible multi-dimensional subspace whose basis vectors each contribute to the model's ability to distinguish propositional true and false statements, revealing a structured internal representation of this kind of truth in language models.

2 Background

2.1 Transformers

Decoder-only transformers (Liu et al., 2018) map input tokens $\mathbf{t} = (t_1, t_2, \ldots, t_n) \in \mathcal{V}^n$ to output probability distributions $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) \in \mathbb{R}^{n\times|\mathcal{V}|}$. Let $\mathbf{x}_i^{(l)}(\mathbf{t}) \in \mathbb{R}^{d_{\text{model}}}$ denote the residual stream activation of the token at position i at the start of layer l.[1] Each token's residual stream is initialized to its embedding $\mathbf{x}_i^{(1)} = \texttt{Embed}(t_i)$, and then undergoes a series of transformations across L layers. Each layer's transformation includes contributions from attention and MLP components:

$$\tilde{\mathbf{x}}_i^{(l)} = \mathbf{x}_i^{(l)} + \texttt{Attn}^{(l)}(\mathbf{x}_{1:i}^{(l)}) \qquad (1)$$

$$\mathbf{x}_i^{(l+1)} = \tilde{\mathbf{x}}_i^{(l)} + \texttt{MLP}^{(l)}(\tilde{\mathbf{x}}_i^{(l)}). \qquad (1)$$

The final logits $\texttt{logits}_i = \texttt{Unembed}(\mathbf{x}_i^{(L+1)}) \in \mathbb{R}^{|\mathcal{V}|}$ are then transformed into probabilities over output tokens $\mathbf{y}_i = \texttt{softmax}(\texttt{logits}_i) \in \mathbb{R}^{|\mathcal{V}|}$.[2]

2.2 Internal Representations of Truth

Recent work suggests that LLMs can encode factuality internally, even if their outputs does not always reflect it (Azaria and Mitchell, 2023). Methods like linear probing and DIM have been used to identify directions in the activation space, often in the residual stream, that correlate with whether a statement is true or false (Marks and Tegmark, 2024; Bürger et al., 2024). We draw inspiration from these works by using labeled data sets of true and false English statements to investigate how the truth is geometrically embedded in the hidden states of the model. Similar to Marks and Tegmark (2024) and Bürger et al. (2024), we define truth as a specific operationalization: simple, unambiguous propositional statements that can be labeled as true or false.

Following Wollschläger et al. (2025), who define refusal properties for vectors, we define analogous truth properties for vectors.

Definition of Truth Property

- *Monotonic Scaling:* when using the direction for activation addition/ablation $\hat{\mathbf{x}}_i^{(l)} = \hat{\mathbf{x}}_i^{(l)} + \alpha \cdot \mathbf{r}$, the model's probability of being more truthful should scale monotonically with α. So, the percentage by which the model flips to the opposite answer (e.g. from no to yes) should scale monotonically with α.

- *Surgical Ablation* Ablating the truth direction through projection

$$\tilde{\mathbf{x}}_i^{(l)} \leftarrow \mathbf{x}_i^{(l)} - \hat{\mathbf{r}}\hat{\mathbf{r}}^{\intercal}\mathbf{x}_i^{(l)}. \qquad (2)$$

should cause the model to shift the answer from an initially true output to a false output.

[1] We shorten $\mathbf{x}_i^{(l)}(\mathbf{t})$ to $\mathbf{x}_i^{(l)}$ when the input $\mathbf{t}$ is clear from context or unimportant.

[2] This high-level description omits details such as positional embeddings and layer normalization.

2.3 Model Interventions

2.3.1 Activation Addition

Given a linear direction vector that represents a concept $\mathbf{r}^{(l)} \in \mathbb{R}^{d_{\text{model}}}$ extracted from layer l, we can use linear interventions such as addition and subtraction, scaled by some coefficient $\alpha \in \mathbb{R}$, to modulate the strength of the corresponding feature in the activation space. For example, adding a learned truth vector to the activations shifts the representation toward regions of the activation space associated with truthful outputs.

$$\mathbf{x}^{(l)\prime} \leftarrow \mathbf{x}^{(l)} + \alpha \cdot \mathbf{r}^{(l)}. \tag{3}$$

Note that for activation addition, we intervene only at layer l, and across all token positions.

2.3.2 Directional Ablation

To investigate the role of a direction $\hat{\mathbf{r}} \in \mathbb{R}^{d_{\text{model}}}$ in the model's computation, we can erase it from the model's representations using *directional ablation* (Arditi et al., 2024).

Directional ablation subtracts the component along $\hat{\mathbf{r}}$ for every residual stream activation $\mathbf{x} \in \mathbb{R}^{d_{\text{model}}}$:

$$\mathbf{x}' \leftarrow \mathbf{x} - \hat{\mathbf{r}}\hat{\mathbf{r}}^{\intercal}\mathbf{x}. \tag{4}$$

We perform this operation at every activation $\mathbf{x}_i^{(l)}$ and $\tilde{\mathbf{x}}_i^{(l)}$, across all layers l and all token positions i. This effectively prevents the model from ever representing this direction in its residual stream.

2.4 Gradient-Based Methods

Gradient-based methods are a class of interpretability techniques that use gradients of model outputs with respect to internal activations to identify influential features or directions by revealing how small changes influence predictions. More recently, Wollschläger et al. (2025) have used gradients to steer model behavior: specific objectives, such as refusing unsafe inputs, can be encoded directly as loss functions. By optimizing a single vector that is added to or ablated from activations at specific layers, models can be guided toward target behaviors (e.g., safe refusals) while minimizing side effects on unrelated outputs. When applied to truthfulness, this framework enables precise, interpretable interventions and allows models to express truth-aligned responses without requiring full fine-tuning.

2.5 Concept Cones

As described in Wollschläger et al. (2025), given a set of orthonormal vectors $V = [v_1, v_2, \ldots, v_k] \in \mathbb{R}^{d_{model} \times k}$ a matrix whose columns are vectors each exhibit truth properties. The cone is the set of all nonnegative linear combinations of

$$\mathcal{R}_N = \{\sum_{i=1}^{k} \lambda_i v_i \mid \lambda_i \geq 0\} \setminus \{0\}$$

All directions used in the cone correspond to the same truth concept. The constraint $\{\lambda_i \geq 0\}$ ensures that all directions within the cone consistently strengthen truth behavior.

3 Methodology

To investigate the existence and structure of directions representing the notion of truthfulness in language models, we start with a linear-probe paradigm introduced by Marks and Tegmark (2024), we locate a linear direction in the residual stream by feeding the model raw factual statements and regressing on their ground-truth labels. We retain their definition of using factual statements that are simple, unambiguous and have topical diversity.

We modify the following: instead of attaching a label offline, we ask the model to answer each statement with a binary "Yes" or "No" and use that forced choice as the supervision signal. This lets us treat the model's own response distribution as a self-labeled probe target, enabling activation addition and ablation tests on the same forward pass and providing targets for our gradient descent approach.

3.1 Setup for Truth Representation Discovery

Each experiment involves prompting the LLM with a short factual statement and requesting a binary "Yes" or "No" response. We format the prompt using a system instruction to make it clear that the model should answer truthfully and concisely (See Appendix C for all system prompts used). For example:

```
System: Respond to the following
statements with either “Yes” or
“No” based on their factual
accuracy.
User: The Eiffel Tower is in
Paris.
Model: Yes ...
```

We assume that for sufficiently capable base models, correct classification is achieved under normal conditions. Our goal is to test whether internal directions in activation space causally mediate this truthful behavior.

3.2 Causal Interventions: Addition and Ablation

Let $\mathbf{r}^{(l)} \in \mathbb{R}^{d_{\text{model}}}$ be a candidate direction vector associated with the concept of truth at layer l. We apply addition and ablation as follows:

- **Directional Addition**: Given a false statement (where the base model typically outputs "No"), we apply $\mathbf{r}^{(l)}$ additively to shift the model's behavior toward "Yes".
- **Directional Ablation**: Given a true statement, we remove the component along $\mathbf{r}^{(l)}$ from the residual stream. If $\mathbf{r}^{(l)}$ encodes truth, the model's output should flip from "Yes" to "No".

3.3 Loss-Guided Concept Cone Discovery

To discover a set of *orthonormal basis vectors* that span a cone encoding the concept of truth, we optimize a composite loss that encourages each vector to:

1. Induce truth behavior when added to false prompts.
2. Inhibit truth behavior when ablated from true prompts.
3. Preserve unrelated model behavior (i.e., maintain fidelity to non-targeted inputs).

Objective. Following Wollschläger et al. (2025), our optimisation target is a three–term loss

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{add}} + \lambda_2 \mathcal{L}_{\text{ablate}} + \lambda_3 \mathcal{L}_{\text{retain}},$$

but with two implementation tweaks that adapt it to binary truth–judgement:

1. **Binary generation.** At generation time we zero out every logit except the two tokens `Yes` and `No` and sample *one* token (t=1), which converts the addition/ablation terms into standard binary cross-entropy losses.
2. **Wide-scope retention.** To guard against collateral drift, $\mathcal{L}_{\text{retain}}$ is measured on 30-token continuations of Alpaca instructions, providing a broad behavioural footprint.

Formally the three components are:

$$\mathcal{L}_{\text{add}} = -\frac{1}{|\mathcal{D}_{\text{false}}|} \sum_{x \in \mathcal{D}_{\text{false}}} \log \hat{y}_{\text{add}}(x + \mathbf{v}) \qquad \text{(Add, target } y{=}1)$$

$$\mathcal{L}_{\text{ablate}} = -\frac{1}{|\mathcal{D}_{\text{true}}|} \sum_{x \in \mathcal{D}_{\text{true}}} \log \left[1 - \hat{y}_{\text{ablate}}(x - \mathbf{v}\mathbf{v}^{\top} x)\right] \qquad \text{(Ablate, target } y{=}0)$$

$$\mathcal{L}_{\text{retain}} = \frac{1}{|\mathcal{D}_{\text{alpaca}}|} \sum_{x \in \mathcal{D}_{\text{alpaca}}} D_{\text{KL}}\big(p_0(y_{1:30} \mid x) \big\| \, p_{\mathbf{v}}(y_{1:30} \mid x)\big) \qquad \text{(Retain, KL)}$$

Here $\hat{y}_{\text{add}}$ and $\hat{y}_{\text{ablate}}$ are the post-softmax probabilities of outputting `Yes` after, respectively, adding or ablating the truth vector $\mathbf{v}$ at the chosen layer; p_0 and $p_{\mathbf{v}}$ denote the unmodified and perturbed 30-token distributions for Alpaca prompts. The scalars $\lambda_{1:3}$ balance steering power ($\mathcal{L}_{\text{add}}, \mathcal{L}_{\text{ablate}}$) against fidelity ($\mathcal{L}_{\text{retain}}$).

4 Experiments

4.1 Experiment 1: Localizing Truth Behavior in Layers and Token Positions

Goal. We investigate which layers and token positions are most effective for capturing truth-related behavior. Since a linear direction is simply a one-dimensional concept cone, we first evaluate whether truth can be causally mediated at each layer using a single direction. If a model fails to encode truth behavior in a linear subspace at a given layer, it is unlikely that a higher-dimensional cone would succeed there either.

Procedure. To do this, we train a one-dimensional cone (i.e., a linear direction) at each layer and across the last five token positions, and evaluate its *Answer Switching Rate* (ASR). Specifically, we measure the success of activation-based interventions across multiple datasets and model families by computing the ASR – the proportion of inputs which affect model outputs after an intervention.

Formally, we define the **Answer Switching Rate (ASR)** as:

Definition of ASR

$$\text{ASR} = \frac{\begin{array}{c}\#\text{ of prompts whose output}\\ \text{becomes untruthful after ablation}\end{array}}{\begin{array}{c}\text{baseline }\#\text{ of prompts that the model}\\ \text{answers truthfully}\end{array}}$$

In practice, the baseline is almost always the same as the total number of prompts as the models

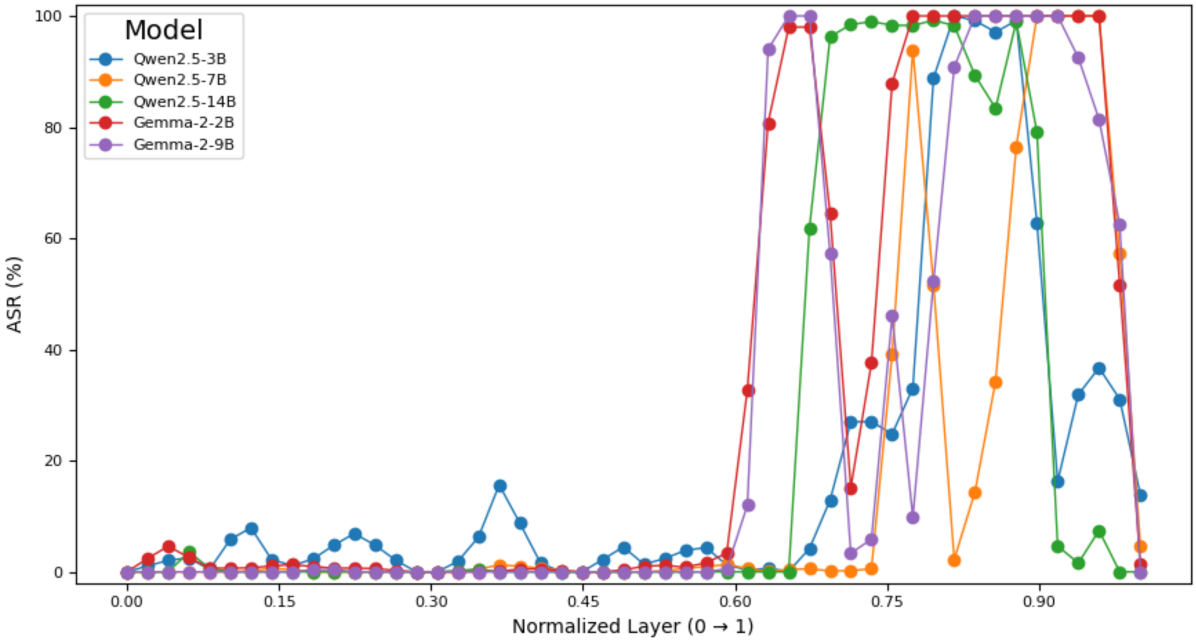

Figure 2: The Answer Switching Rate (ASR) of one dimensional cones across layers for Qwen and Gemma models. The layer numbers have been normalized across larger and smaller models. The effectiveness spikes rapidly in all models in the 0.60-0.75 range of normalized layer numbers.

nearly always achieve full accuracy when answering our simple propositions.

Results. Across both model families and sizes, we find that truth-related directions reliably emerge in the middle to later layers (specifically, between 60–75 percent of the normalized layer depth). As shown in Figure 2, ASR increases sharply in this range before decreasing sharply again in the very last layers. Additionally, we find that the final token position consistently yields the strongest interventions, consistent with prior work showing that high-level decisions often accumulate at the end of the sequence (Arditi et al., 2024; Bürger et al., 2024).

Based on these findings, we restrict our concept cone search to this high-performing region of the network. This choice is motivated both by empirical signal strength and by computational efficiency.

4.2 Experiment 2: Truthfulness Steering Across Models and Dimensions with Cones

Goal. The aim of this experiment is to assess the effect of increasing dimensionality on the ability causally mediate truthful behavior. While previous results show that a single direction can causally influence truthfulness, we seek to determine how many additional, orthogonal directions can also support this behavior before unrelated features begin to dilute the effect. We do this across multiple models from the Qwen-2.5 and Gemma-2 families, spanning a range of parameter sizes.

Procedure. For each model, we construct concept cones with dimensionalities ranging from 1 to 5. Each cone is generated using the optimization procedure described in Section 3, which ensures the basis vectors satisfy both causal and retention constraints. To evaluate generalization across the cone space, we perform Monte Carlo sampling within each cone by drawing random nonnegative combinations of the basis vectors. We then measure the effectiveness of each sampled direction using ASR defined previously.

Results Table 1 presents the Answer Switching Rate (ASR) across five language models (Qwen2.5-3B, Qwen2.5-7B, Qwen2.5-14B, Gemma-2-2B, and Gemma-2-9B) as a function of the dimensionality of the concept cone used for intervention. Each ASR value reflects the average success rate across Monte Carlo samples drawn from the cone of that dimension.

Interpretation. The results in Table 1 suggest that increasing the dimensionality of the concept cone generally improves the model's ability to internalize and respond to truth-aligned interventions.

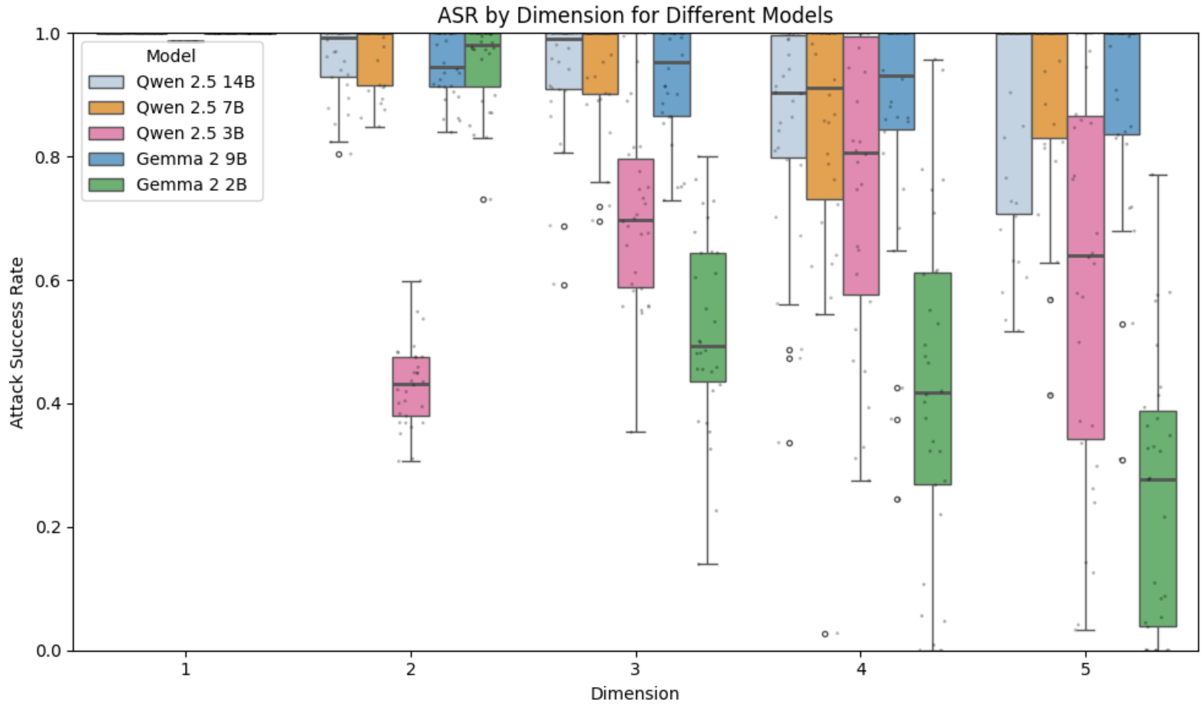

Figure 3: The Answer Switching Rate (ASR) of cones from dimensions 1 to 5 across Qwen2.5 and Gemma2 models with boxplots showing the Monte Carlo sampling.

3

Table 1: Answer Switching Rate after intervention across models and cone dimensions.

Model	1(DIM)	2	3	4	5
Qwen 14B	100	100	98.6	91.2	100
Qwen 7B	100	100	100	100	100
Qwen 3B	98.6	45.1	67.2	78.9	65.3
Gemma 9B	100	100	100	98.6	97.3
Gemma 2B	100	100	53.7	43.1	27.1

Larger models, such as Qwen-2.5-7B and Gemma-2-9B, maintain high ASR even as dimensionality increases, meaning that higher dimension cones exist within their activation space. This is consistent with findings from Wollschläger et al. (2025) in the domain of refusal behavior.

However, the trend is not monotonic: beyond a certain point, ASR begins to decline, indicating that additional directions may start to capture unrelated features and dilute the effectiveness of the intervention. This effect is especially evident in smaller models, where cone dimensions above 2 or 3 yield diminishing or negative returns. Nonetheless, the models still show multiple dimensions that independently support truth-aligned behavior. For example, both Qwen-7B and Gemma-9B maintain near-100% ASR across all tested dimensions, showing that there is at least a 5-dimensional cone that causally mediates truth.

4.3 Experiment 3: Retention of General Capabilities via KL Divergence

Goal. While the purpose of truth-direction interventions is to modify the model's factual response behavior, we must ensure they do not interfere with unrelated capabilities. This experiment evaluates how much the intervention alters model output on a general instruction-following benchmark, using KL divergence as a metric of deviation. This operationalizes the $\mathcal{L}_{\text{retain}}$ loss term defined in Section 3

Dataset. We use the *ALPACA* (Taori et al., 2023) dataset, a popular instruction-following benchmark designed to elicit helpful, safe, and general-purpose completions. We randomly select 200 prompts that are unlikely to invoke factual disputes (e.g., summarization, rewriting, math, or basic instructions).

Procedure. We use the *ALPACA* (Taori et al., 2023) test set, a popular instruction-following benchmark designed to elicit helpful, safe, and general-purpose completions. We randomly select 200 prompts that are unlikely to invoke factual disputes (e.g., summarization, rewriting, math, or basic instructions).

For each cone that we generate, we compare the original model's outputs to those produced after applying directional ablation using the discovered truth directions. The intervention is applied globally (all tokens, all layers) as described in Equation 4. As a threshold, we don't consider cones with basis vectors with a KL Divergence above 0.1, as Following the precedent of papers like (Arditi et al., 2024), we use 0.1 as a threshold value for the KL consideration of .

Results. We report the mean KL divergence across 200 Alpaca prompts in Table 2. We find that the truth-direction ablation leads to only minimal divergence from the original output distribution, suggesting that the intervention does not significantly affect unrelated capabilities.

Table 2: Mean KL divergence on Alpaca prompts (lower is better).

Model	Mean KL Divergence
Qwen2.5-14B	0.038
Gemma-2-2B	0.045
Qwen2-7B	0.026
Gemma-2-9B	0.031

Interpretation. All models show low average KL divergence, especially the larger variants. This suggests that the discovered truth directions are highly specific and do not interfere with general instruction-following behavior. The effectiveness of $\mathcal{L}_{\text{retain}}$ as a regularization objective is empirically supported by this result.

4.4 Experiment 4: DIM vs. Cone Alignment

We measure how closely the classic *Difference-in-Means* (DIM) truth vector aligns with the orthonormal directions discovered by our Concept Cone. Cosine similarity is reported in Table 3; values near 1 indicate strong overlap.

Table 3: Cosine similarities between the DIM direction and cone basis vectors in Gemma-2-9B, transposed for dimensions 2–5.

	Dim 2	**Dim 3**	**Dim 4**	**Dim 5**
v_1	1.23×10^{-1}	1.45×10^{-1}	2.00×10^{-1}	2.26×10^{-1}
v_2	-3.72×10^{-9}	1.74×10^{-9}	3.03×10^{-9}	-6.98×10^{-10}
v_3	—	1.16×10^{-9}	-2.33×10^{-9}	-4.19×10^{-9}
v_4	—	—	2.33×10^{-10}	8.38×10^{-9}
v_5	—	—	—	3.03×10^{-9}

Results. Only the first cone axis has any alignment with DIM, confirming that DIM captures just one facet of the multi-dimensional truth subspace; the remaining axes encode additional, orthogonal structure.

5 Discussion

Our findings reveal that while a single direction derived from the DIM method already captures a strong causal representation of truth in LLMs, it does not fully exhaust the structure underlying truth-related behavior. Through our concept cone approach, we identified additional orthogonal directions with low cosine similarity to the DIM vector that also reliably steer model outputs on propositional truth tasks. This suggests that truthful behavior may not be confined to a single axis—multiple directions and can be independently influenced. These directions likely correspond to distinct or semantically adjacent components of factual reasoning, such as modality, certainty, or domain-specific features.

The success of both DIM and cone-based interventions suggests that truth may be linearly separable in the model's representation space. While the directions that independently modulate truthful behavior can imply that this structure may be richer than a single linear axis, it does not necessarily prove that the underlying representation of truth is nonlinear. It does, however, open up important questions in the context of model deception, robustness, and interpretability. If multiple, semantically adjacent directions can influence truthfulness, models may be more vulnerable to manipulations that subtly shift their outputs without obvious signs of tampering within the first truth direction. Understanding the geometry of these truth-related subspaces is essential for building models that are not only aligned, but resilient to adversarial or unintended shifts in behavior.

6 Conclusion and Future Directions

In this work, we showed that multi-dimensional concept cones can reliably steer the behavior of LLMs on True/False propositions across multiple architectures and sizes, while minimally impacting unrelated behaviors. Our results reveal that, beyond a single "truth direction," there exists a robust subspace of activation vectors whose positive combinations consistently modulate factuality. These findings show increasing promise for concept cones as an interpretability toolkit and underscore new avenues and risks for alignment, calibration, and adversarial manipulation of model truthfulness.

Several promising avenues remain. Concept-cone search reliably uncovers a subspace, but we still are yet to find semantically meaningful labels for the basis vectors. Future work could pair cones with automated clustering or sparse coding so that each basis vector corresponds to an interpretable facet of truth (e.g., temporal facts, geographic facts, commonsense). Extending the method to larger, instruction-tuned models and to multimodal settings would also test its robustness and reveal whether these semantic dimensions persist across scale and modality.

7 Limitations

7.1 Model Scale

All experiments were conducted on relatively small open-source models (1.5B–7B parameters). While we observe clear directional structure in the residual stream of mid-sized models, these findings may not generalize to larger frontier models or architectures with substantially different alignment protocols. Notably, smaller models exhibit lower ASR, with PCA visualizations revealing weaker separation of truth-related directions, especially in later layers. This suggests that representational abstraction of truth may emerge more clearly with scale.

7.2 Scope

Our operationalization of truth/factfulness is deliberately narrow, limited to simple unambiguous propositional facts that have a clear true or false answer. While this allowed for clean experimental design, it does not capture more complex notions of truth that may be more widely applicable, such as context-dependent claims or statements that involve some kind of subjectivity. As a result, the discovered directions or cones may not generalize to broader or more nuanced conceptions of truth. Future work should explore whether similar geometric structures exist for more complex truth representations, and whether the cone framework can be extended to handle contextual, graded, or higher-order reasoning tasks.

We also restrict model outputs to binary "Yes"/"No" responses via logit masking. While ensuring clarity in supervision and evaluation, it artificially simplifies generation task and limit applicability to more open-ended settings. In a more realistic setting, models express uncertainty and nuanced responses that binary outputs do not capture.

Our experiments span only two model families (Gemma and Qwen). It remains an open question whether the discovered directions are robust to architectural variation or fine-tuning differences. Evaluating cross-family generalization, especially to models trained with stronger alignment (e.g., RLHF or human preference tuning), is an important direction for future work.

7.3 Subspace Understanding

Although we demonstrate that a low-dimensional subspace (or “cone”) can causally mediate truth behavior, our method does not guarantee the discovery of a maximally informative or interpretable subspace. We leave for future work the development of principled methods for example using sparsity constraints, disentanglement metrics, or unsupervised clustering—to assign semantic meaning to individual cone axes.

References

Sotiris Anagnostidis and Jannis Bulian. 2024. How susceptible are llms to influence in prompts? *arXiv preprint arXiv:2408.11865*.

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

Amos Azaria and Tom M. Mitchell. 2023. The internal state of an llm knows when it’s lying. *Preprint*, arXiv:2304.13734.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety – a review. *arXiv preprint arXiv:2404.14082*.

Tom Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in llms. *arXiv preprint arXiv:2407.12831*.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, and 2 others. 2024. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*, pages 2254–2272. ACM.

Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, Xu Shen, and Jieping Ye. 2024. From yes-men to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning. *arXiv preprint arXiv:2409.01658*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Joshua Engels, Isaac Liao, Eric J Michaud, Wes Gurnee, and Max Tegmark. 2024. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *the Twelfth International Conference on Learning Representations*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Dan Hendrycks and Mantas Mazeika. 2022. X-risk analysis for ai research. *CoRR*, abs/2206.05862.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

Jaewoo Heo, Christoph Heinze-Deml, Omar Elachqar, Shuning Ren, Udaya Nallasamy, Andrew Miller, Ka Ho Roy Chan, and Janarthanan Narain. 2025. Do llms "know" internally when they follow instructions? In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Fabian Hildebrandt, Andreas Maier, Patrick Krauss, and Achim Schilling. 2025. Refusal behavior in large language models: A nonlinear perspective. *arXiv preprint arXiv:2501.08145*.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. 2023. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth Daly, and Kush R. *et al.* Varshney. 2025. Evaluating the prompt steerability of large language models. In *Proceedings of NAACL 2025*, pages 7874–7900.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Richard Ngo, Lawrence Chan, and Soren Mindermann. 2022. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt/`. Accessed: 2025-01-26.

OpenAI. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024a. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024b. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *International Conference on Machine Learning*, pages 39643–39666. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, and 1 others. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Lee Sharkey, Dan Braun, and Beren Millidge. 2023. Taking features out of superposition with sparse autoencoders. Accessed: 2025-04-22.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Rishi Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://`

github.com/tatsu-lab/stanford_alpaca. Accessed: 2025-05-17.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear Representations of Sentiment in Large Language Models. *arXiv preprint arXiv:2310.15154*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Haoran Wang and Kai Shu. 2023. Trojan activation attack: Red-teaming large language models using activation steering for safety-alignment. *arXiv preprint arXiv:2311.09433*.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Yuan Zhou, Meng Liu, and Xinyu Li. 2025. Understanding the relationship between prompts and response uncertainty in large language models. In *ICLR 2025 Workshop on Building Trustworthy AI*.

A Setup Details

A.1 Implementation Details

Experiments run on NVIDIA H100-80GB GPUs using `PyTorch` 2.20 and `HF Transformers` 4.41. Core settings are in Table 6.

Table 4: Hardware and hyper-parameters.

GPUs	1×H100 (probing)
Batch size	4
Number of Samples (during training)	16
Precision	bfloat16
Optimizer	AdamW

Code base	`transformer-lens 0.9.1`
	`nnsight 0.3.7`

A.2 Datasets

Table 5: Full datasets used in experiments

Cities	Link
Animals	Link
Elements	Link

A.3 Models

Table 6: All Models used in Experiments

Qwen2.5-3B-Instruct	Link
Qwen2.5-7B-Instruct	Link
Qwen2.5-14B-Instruct	Link
Gemma-2-2B-IT	Link
Gemma-2-9B-IT	Link

B Additional Experiments

B.1 Sentiment

Previous literature (Tigges et al., 2023) suggests that sentiment has a linear representation, similar to other concepts such as refusal (Arditi et al., 2024). We tried to extend our methodology to sentiment to determine whether it has a concept cone representation. In particular, we trained a concept cone on the Stanford Sentiment Treebank (Socher et al., 2013) which consists of 10,662 one-sentence movie reviews with fully labeled parse trees. We failed to find a meaningful concept cone for sentiment. Further work could explore alternative techniques for finding a higher-dimensional representation for sentiment.

B.2 Toxicity

We explored the existence of higher-dimensional representations for toxicity. We trained a concept cone on the ToxiGen benchmark (Hartvigsen et al., 2022) which consists of 274,000 toxic and benign phrases about minority groups such as such as Black, Asian, Latino, Jewish, Muslim, LGBTQ+, women, and individuals with mental or physical disabilities. We first computed a difference-in-means (DIM) direction. We also computed a direction by taking the difference between the activations of a student model that generates toxic output (typically an instruction-tuned model) and a teacher model that doesn't generate toxic output (typically an instruction-tuned model) on the dataset, a method described in Wang and Shu (2023). Testing this direction using ablation failed, as the resulting output was unintelligible. Testing the DIM direction using ablation did disable strong refusal (the model did not say "I cannot respond" outright) but would still refuse to generate toxic output. Since we were unable to obtain a valid linear direction for toxicity, we were unable to generate high-quality targets and as a result, were unable to train a valid cone. Further work could explore better methods for obtaining a direction for toxicity.

C Full List of System Prompts

System-level instructions exert a strong influence over an LLM's eventual answer: recent work shows that even minor wording differences can sway a model toward a different multiple-choice option, particularly when those differences encode authority or confidence (Anagnostidis and Bulian, 2024); that prompts which explicitly acknowledge uncertainty often elicit more calibrated responses (Zhou et al., 2025); and that a model's overall "steerability" rises or falls with the directness of the request phrasing (Miehling et al., 2025).

Table 7: Example system prompts that force a binary `Yes`/`No` answer.

#	**Prompt wording**
1	Respond to the following statement with a **"Yes"** or a **"No."**
2	Respond to the following statements with either **"Yes"** or **"No"** based on their factual accuracy.
3	Answer with **exactly one word—"Yes"** if the statement is true, otherwise **"No"**.
4	Based on your best estimate of the following statement, respond **"Yes"** or **"No."**

D Token choice to represent True/False

In our experiments, we frame factual statements as binary questions and constrain the model's output to the tokens `Yes` and `No`. We also tested alternative tokenizations such as `yes/no` and `true/false`, but found that these variations had no significant effect on steering outcomes or ASR. Interestingly, when the output vocabulary is left unrestricted, models occasionally respond in non-English equivalents of "Yes" and "No" (e.g., `Sí`, `Nein`) following truth-direction interventions. This suggests that the underlying truth representation may generalize across lexical choices, although further investigation is needed to confirm cross-lingual consistency.

E Cosine Similarities

We see the same trend for cosine similarity across models, where other than the first dimension, all increasing dimensions have extremely low cosine similarity to the DIM direction.

Table 8: Cosine similarities between the DIM direction and cone basis vectors in Qwen-2.5-9B, transposed for dimensions 2–5.

	Dim 2	**Dim 3**	**Dim 4**	**Dim 5**
v_1	-1.57×10^{-1}	1.82×10^{-1}	1.34×10^{-1}	1.67×10^{-1}
v_2	-4.23×10^{-9}	2.91×10^{-9}	-1.08×10^{-9}	-5.74×10^{-10}
v_3	—	3.56×10^{-9}	-7.42×10^{-9}	6.13×10^{-9}
v_4	—	—	2.87×10^{-9}	-2.45×10^{-10}
v_5	—	—	—	4.39×10^{-9}

F Code

All code will be open-sourced on Github.

G PCA Visualizations

As a cursory introduction into understanding literature of linear representations of truth, we recreated Principal Component Analysis visualizations of all models used in the experiments on the datasets onto their top two principal components. All components are listed below.

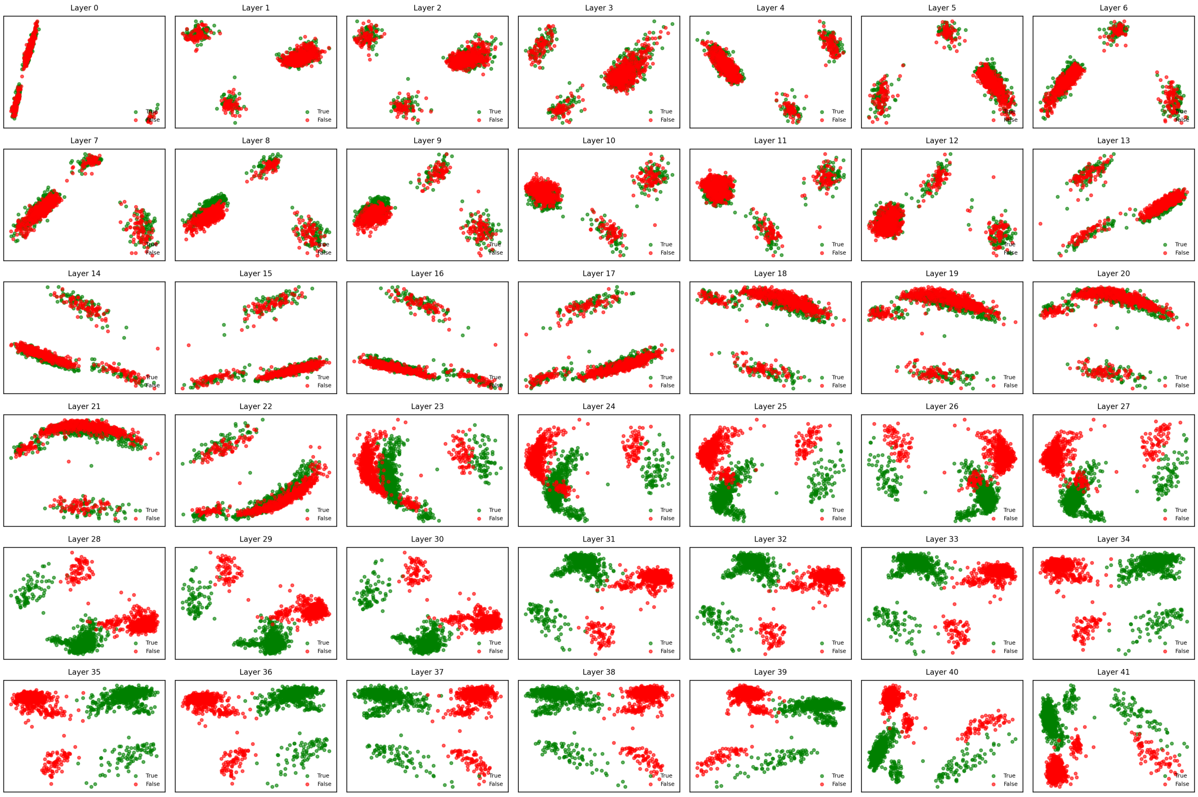

Figure 4: Projections of Gemma-2-9B, representations of datasets onto their top two PCs, across all layers.

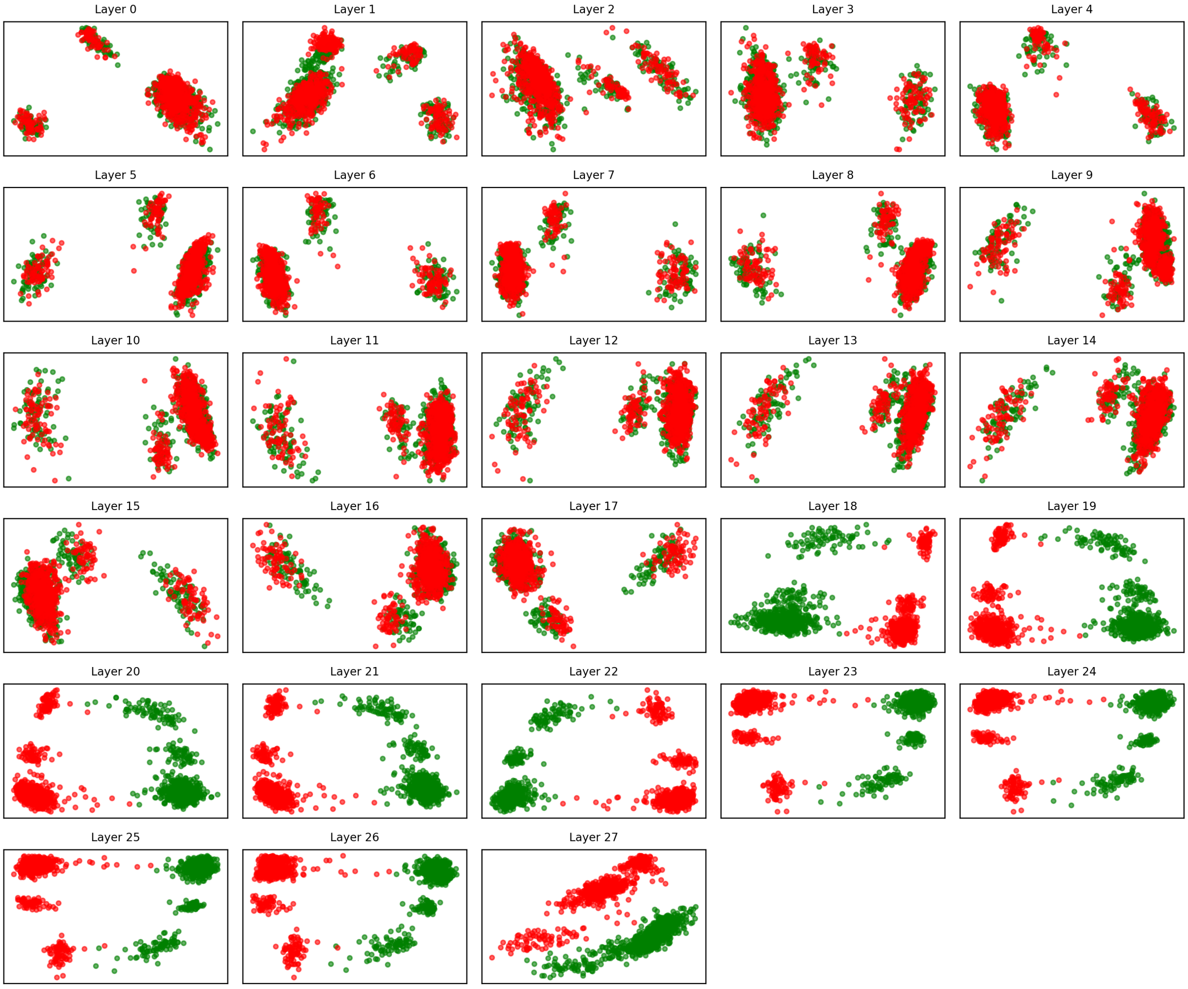

Figure 5: Projections of Qwen2.5-7B representations of datasets onto their top two PCs, across all layers.