Spectral Principal Paths: A Spectral Perspective on Linear Representation Formation in LLMs

Bowei Tian[1] **Xuntao Lyu**[2] **Meng Liu**[1] **Hongyi Wang**[3] **Ang Li**[1]

[1]University of Maryland, College Park
[2]North Carolina State University
[3]Genbio AI

Abstract

High-level representations have become a central focus in enhancing AI transparency and control, shifting attention from individual neurons or circuits to structured semantic directions that align with human-interpretable concepts. While the Linear Representation Hypothesis (LRH) suggests that such directions emerge in representations, it remains unclear how these representations originate and why they become increasingly stable across layers. To solve this issue, we introduce the Input-Space Linearity Hypothesis, positing that concept-aligned directions originate in the input space and are steadily maintained with increasing depth. We then propose the Spectral Principal Path (SPP) framework, which formalizes how deep networks progressively distill linear representations along the spectral principal directions. We provide rigorous stability guarantees for the SPP based on the Wedin $\sin\Theta$ perturbation theorem, identifying testable conditions, including spectral gap and context incoherence, that jointly ensure layer-wise directional preservation. By bridging theoretical analysis with empirical evidence, this work identifies a spectral view of how linear representations arise in LLMs, and suggests potential implications for concept-level controllable, robust, and coherent approaches to fairness and transparency in modern AI systems.

1 Introduction

Large Language Models (LLMs) have demonstrated strong performance across a wide range of language-related tasks, including question answering (Brown et al., 2020; Zou et al., 2023), code generation (Chen et al., 2021), and reasoning (Wei et al., 2022). However, the internal mechanisms of neural networks remain opaque. Despite advances in interpretability techniques, how LLMs transform inputs into high-level representations remains poorly understood (Nanda et al., 2023; Weidinger et al., 2021). This opacity raises concerns about reliability in high-stakes domains such as healthcare, finance, and law (Bommasani et al., 2021; Weidinger et al., 2021; Conmy et al., 2023).

Previous works have demonstrated the potential of representations as a new perspective on LLM transparency. For example, LLMs have been shown to internally encode linguistic structure across layers, with different layers capturing syntax and semantics at varying levels of abstraction (Devlin et al., 2019; Belinkov, 2022). More recently, structured representations of world knowledge, including spatial and temporal information, have been found to emerge spontaneously in language model hidden states (Gurnee & Tegmark, 2023; Meng et al., 2022). Zou et al. (2023) further formalized Representation Engineering (RepE), emphasizing its ability to extract meaningful concepts from a model's internal structure and control model behavior. RepE has emerged as a top-down approach to enhance model transparency that focuses on representations rather than individual neurons or circuits, providing a more structured understanding of AI transparency and control. Another important contribution is the Linear Representation Hypothesis (LRH) (Park et al., 2023): as depth increases, task-

relevant concepts become nearly linearly separable in the model's latent space, making them accessible with simple probes or linear edits.

Despite these promising advances, a theoretical understanding of how structured representations emerge and evolve across layers in LLMs remains to be fully established. While approaches such as RepE and LRH have offered valuable insights into the geometry of concept directions, questions remain about why representations become linearly organized and how they propagate through LLMs. A more complete account has yet to emerge — in particular, it remains unclear whether the linear structure observed in LLM representations is an intrinsic property of the model's learned parameters, or whether it reflects structure already present in the input space prior to any learned transformation.

In this work, we propose Spectral Principal Paths (SPP) as a theoretical framework for investigating the emergence and stability of linear representations in LLMs. We introduce the Input-Space Linearity Hypothesis (ISLH), which suggests that concept-aligned directions originate in the input space and are maintained with increasing depth. These representations propagate through the spectral principal paths, and empirical evidence is in support of this view. This structure offers a potential explanation for why concept directions remain relatively stable and linearly accessible across layers, and sheds light on the theoretical underpinnings of RepE and LRH. We hope this framework contributes a step toward bridging theory and practice in understanding representations in LLMs. Code is released at https://github.com/boweitian/SPP.

Our main contributions are as follows:

- **Input-Space Linearity Hypothesis (ISLH).** We extend the Linear Representation Hypothesis beyond the representation space to the input space, showing that concept directions can be traced backward to the input space.
- **Spectral Principal Path (SPP) with rigorous stability guarantees.** We propose the SPP mechanism and provide a formal stability analysis based on the Wedin $\sin\Theta$ perturbation theorem. We identify testable conditions, including spectral gap and context incoherence, under which the SPP direction is provably preserved across layers, with quantitative bounds on angular deviation.
- **Systematic Empirical Validation and Causal Analysis.** We validate the SPP framework across four LLMs, confirming that spectral structures remain robust across structures. Through controlled causal ablations, we provide evidence that both spectral gap and context incoherence are necessary conditions for the emergence and stability of linear concept representations in transformer networks.

2 Related Works

2.1 Representation Learning

Early work on word embeddings shows that neural networks can learn distributed representations that encode semantic relationships and compositional structures (Mikolov et al., 2013). Follow-up studies (Schramowski et al., 2019; Radford et al., 2015) further reveal that learned embeddings can implicitly encode abstract dimensions such as commonsense morality, even without explicit supervision. For instance, Radford et al. (Radford et al., 2015) observe that training a language model on product reviews results in the emergence of a sentiment-tracking neuron.

This phenomenon is not unique to language models. McGrath et al. (2022) show that similar internal representations can be found in networks trained to play chess. In computer vision, recent studies (Caron et al., 2021; Oquab et al., 2023) demonstrate that both generative and self-supervised training objectives give rise to emergent semantic representations, such as those useful for segmentation tasks, suggesting the emergence of representations is a general property of deep learning systems.

Building on this, Zou et al. (2023) propose techniques to read and control these internal structures, including Linear Artificial Tomography (LAT) for extracting concept-aligned

representations and methods for steering model behavior. Their study shows that RepE-style approaches can be used not only to detect but also to manipulate emergent properties, motivating more systematic efforts to characterize and intervene in high-level model behaviors. Theoretically, Park et al. (2023) proposes the Linear Representation Hypothesis: task-relevant concepts become nearly linearly separable in the model's latent space.

2.2 Approaches to Interpretability

Traditional interpretability techniques have focused on methods such as saliency maps (Simonyan et al., 2013; Springenberg et al., 2014; Zeiler & Fergus, 2014; Zhou et al., 2016), feature visualization (Szegedy et al., 2013; Zeiler & Fergus, 2014), and mechanistic interpretability (Olah et al., 2020; Olsson et al., 2022; Lieberum et al., 2023). These approaches provide useful local explanations by attributing predictions to specific regions, neurons, or circuits. Saliency maps (Simonyan et al., 2013) identify influential input regions via gradients or activations, but they are often unstable and offer limited insight into how information is distributed across layers. Similarly, feature visualization (Szegedy et al., 2013; Zeiler & Fergus, 2014) targets individual neurons, which may overlook the broader structure of learned representations. Mechanistic interpretability (Olah et al., 2020; Olsson et al., 2022; Lieberum et al., 2023; Zou et al., 2023) aims to reverse engineer networks into interpretable circuits or algorithmic components. While it has yielded compelling case studies, this approach typically requires substantial manual effort and remains difficult to scale to large models. Moreover, the assumption that complex neural computations can be cleanly decomposed into discrete circuits is theoretically challenging, given the highly distributed and high-dimensional nature of modern representations.

In response to these limitations, more recent work has shifted attention toward the geometry and dynamics of representation spaces. Rather than attributing behavior to individual neurons or circuits, these approaches study how entire representation structures evolve across layers and models. Techniques such as CKA (Kornblith et al., 2019), SVCCA (Raghu et al., 2017), and Procrustes alignment (Gao et al., 2021) quantify the similarity and transformation of representations, revealing systematic layer-wise reorganization and shared subspaces across architectures. Complementary lines of research examine how architectural features shape representational flow: residual connections have been shown to support stable feature reuse and smoother optimization trajectories (He et al., 2016; Dehghani et al., 2023). However, these studies largely characterize representational geometry in a descriptive manner, without explaining how such geometric structures are formally propagated, preserved, or amplified across layers.

3 Preliminaries

Linear Representation Hypothesis (Park et al., 2023) We consider a concept $W \in \{0, 1\}$ that has a *linear representation* in a model if there exists a vector $\bar{\gamma}_W$ in the unembedding space Γ and a vector $\bar{\lambda}_W$ in the embedding space Λ such that for any counterfactual pair $(Y(W = 0), Y(W = 1))$,

$$\gamma(Y(W = 1)) - \gamma(Y(W = 0)) \in \text{Cone}(\bar{\gamma}_W), \tag{1}$$

and for any context pair (λ_0, λ_1) that changes only W and not other causally separable concepts,

$$\lambda_1 - \lambda_0 \in \text{Cone}(\bar{\lambda}_W). \tag{2}$$

where $\text{Cone}(\mathbf{v}) = \{\alpha\mathbf{v} \ : \ \alpha > 0\}$, the embedding space is where input contexts are mapped to high-dimensional vectors before processing, capturing the model's internal representation of the input. The unembedding space is where each output token is represented, and predictions are made by computing inner products between input embeddings and output unembedding vectors. Unless stated otherwise, all discussions pertain to the embedding space Λ, as our goal is to trace how input linearity propagates through the network.

In the remainder of the paper, we adopt the symmetric encoding $c \in \{-1, 1\}$ for binary concepts, which is equivalent to $c = 2W - 1$.

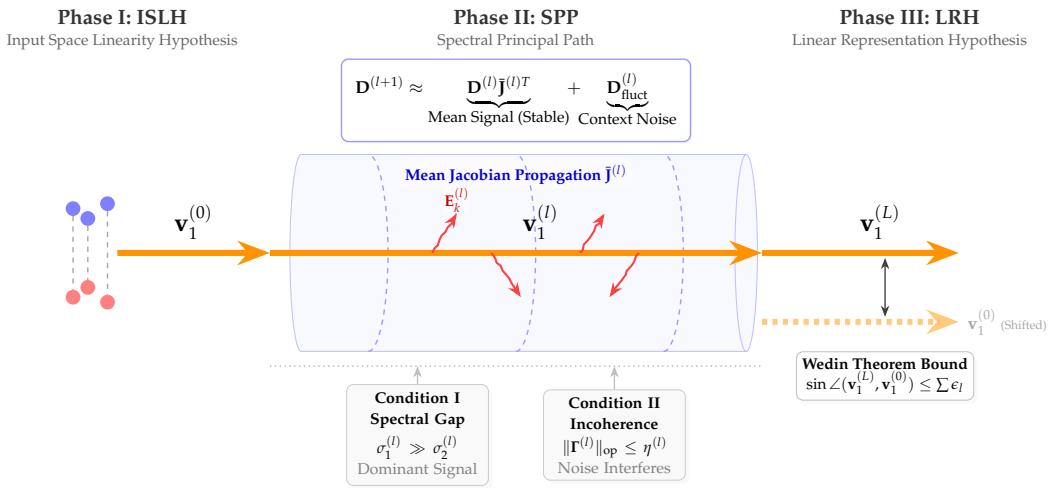


Figure 1: Overview of the **Spectral Principal Path** framework. The conceptual seed $\mathbf{v}_1^{(0)}$ originates in the input space (Phase I: ISLH) and is robustly propagated through network layers via the mean Jacobian $\bar{\mathbf{J}}^{(l)}$ (Phase II). The stability of this path is mathematically guaranteed by (I) Spectral Gap and (II) Context Incoherence, which collectively suppress the destructive influence of context fluctuations. Ultimately, the cumulative angular drift is strictly bounded, culminating in the stable linear representations observed in deep layers (Phase III: LRH).

4 Formalizing the Emergence of Linear Representations

In this section, we provide a formal mathematical framework demonstrating how the Linear Representation Hypothesis (LRH) is naturally derived from the Input-Space Linearity Hypothesis (ISLH) via the formation of highly stable Spectral Principal Paths (SPP). The overview of the Spectral Principal Path framework is shown in Fig. 1. We first define the key objects (§4.1–4.3), then state the formal stability guarantee (§4.4), and finally derive LRH from ISLH (§4.5).

4.1 Input-Space Linearity Hypothesis (ISLH)

Let $X \in \mathbb{R}^{N \times d}$ denote the input sequence of length N with hidden dimension d. We focus on the representation of the last token x_N, which is typically used for sequence-level predictions.

Definition 4.1 (Input-Space Linearity Hypothesis)**.** Let $c \in \{-1, 1\}$ denote a binary semantic concept (e.g., honest vs. dishonest). The ISLH posits that the linear structure described by LRH is a concept-aligned direction already present in the input embedding space:

$$x_i^{(c)} = \mu_i + c \cdot \alpha_i v_{in} + \epsilon_i, \quad \forall i \in \{1, \ldots, N\} \tag{3}$$

where $\mu_i \in \mathbb{R}^d$ is the concept-agnostic base semantics of token i, $\alpha_i > 0$ is the token-specific expression strength of the concept, $v_{in} \in \mathbb{R}^d$ is the unit concept direction in the input space, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$ is orthogonal noise. Under this view, the model's role is not to create such directions, but to selectively preserve them across layers via spectral principal paths, which serve as the primary channels through which input-space structure propagates to linear representations.

4.2 Information Aggregation via Attention

Initially, the linear concept direction v_{in} is distributed across all N tokens. We first formalize how the attention mechanism in the shallowest layers aggregates this distributed signal into the last token x_N.

Lemma 4.2 (Attention-Driven Signal Concentration). *Let $h_1^{(N)} \in \mathbb{R}^d$ be the hidden state of the last token after the first Transformer block. Under the residual attention mechanism, the concept signal is linearly aggregated and amplified:*

$$\Delta h_1^{(N)} = 2\left(\alpha_N I + W_V \sum_{i=1}^{N} A_{N,i}\alpha_i\right) v_{in} := \tilde{v}_1 \tag{4}$$

where $\Delta h_1^{(N)} = \mathbb{E}[h_1^{(N)}|c=1] - \mathbb{E}[h_1^{(N)}|c=-1]$ is the concept difference and $A_{N,i}$ are the softmax attention weights.

The proof follows by substituting the ISLH formulation into the residual attention update $h_1^{(N)} = x_N + \sum_{i=1}^{N} A_{N,i} W_V x_i$ (see Section A.1). Thus, the self-attention mechanism acts as a spatial pooling operator, folding the distributed input-space concept v_{in} into a single dominant concentrated direction $\tilde{v}_1$ at the last token position.

4.3 Spectral Principal Path

Having established that the concept signal concentrates into a direction $\tilde{v}_1$ at the last token after the first layer, we now formalize how this direction propagates through subsequent layers. The main idea is to study the *population-level* behavior across multiple probing sequences (i.e., paired input sequences designed to elicit contrastive concept expressions), rather than a single input.

Definition 4.3 (Difference Matrix and SPP). Given M paired sequences $\{(s_k^+, s_k^-)\}_{k=1}^{M}$ differing only in *concept tokens* (i.e., the tokens whose content determines the concept label c), let $\mathbf{h}_k^{+,(l)}, \mathbf{h}_k^{-,(l)} \in \mathbb{R}^d$ denote the last-token hidden state at layer l. Define the *concept difference vector* $\mathbf{d}_k^{(l)} = \mathbf{h}_k^{+,(l)} - \mathbf{h}_k^{-,(l)}$ and the *difference matrix*:

$$\mathbf{D}^{(l)} = \begin{pmatrix} (\mathbf{d}_1^{(l)})^\top \\ \vdots \\ (\mathbf{d}_M^{(l)})^\top \end{pmatrix} \in \mathbb{R}^{M\times d}. \tag{5}$$

The *Spectral Principal Path* (SPP) at layer l is the leading right singular vector $\mathbf{v}_1^{(l)} \in \mathbb{R}^d$ of $\mathbf{D}^{(l)}$, corresponding to the largest singular value $\sigma_1^{(l)}$.

Definition 4.4 (SPP Stability). The SPP is *ϵ-stable* between layers l and $l+1$ if $|\langle \mathbf{v}_1^{(l)}, \mathbf{v}_1^{(l+1)}\rangle| \geq 1-\epsilon$.

Layer-wise dynamics. For the k-th sequence, the last-token hidden state $\mathbf{h}_k^{(l)} \in \mathbb{R}^d$ evolves through a nonlinear map whose Jacobian we denote $\mathbf{J}_k^{(l)} = \partial \mathbf{h}_k^{(l+1)} / \partial \mathbf{h}_k^{(l)} \in \mathbb{R}^{d\times d}$. Due to residual connections, $\mathbf{J}_k^{(l)} = \mathbf{I} + \mathbf{\Delta}_k^{(l)}$, where $\mathbf{\Delta}_k^{(l)}$ aggregates the attention and MLP gradients. To first order:

$$\mathbf{d}_k^{(l+1)} \approx \mathbf{J}_k^{(l)} \mathbf{d}_k^{(l)} + \boldsymbol{\xi}_k^{(l)}, \tag{6}$$

where $\boldsymbol{\xi}_k^{(l)}$ is a higher-order remainder. We decompose the Jacobian perturbation across sequences into a population mean and individual fluctuation:

$$\mathbf{\Delta}_k^{(l)} = \bar{\mathbf{\Delta}}^{(l)} + \mathbf{E}_k^{(l)}, \quad \text{where } \bar{\mathbf{\Delta}}^{(l)} = \frac{1}{M}\sum_{k=1}^{M} \mathbf{\Delta}_k^{(l)}. \tag{7}$$

Here $\mathbf{E}_k^{(l)} \in \mathbb{R}^{d\times d}$ is the context-specific Jacobian fluctuation for the k-th sequence, capturing how sequence k's Jacobian deviates from the population mean.

4.4 Stability of the Spectral Principal Path

We now state the formal conditions under which SPP stability is guaranteed. Each assumption is an empirically testable condition rather than a structural consequence of the architecture.

Condition 1 (Spectral Gap). *At each layer l, the difference matrix $\mathbf{D}^{(l)}$ has a nontrivial spectral gap: $\sigma_1^{(l)} \gg \sigma_2^{(l)}$, where $\sigma_1^{(l)}$ and $\sigma_2^{(l)}$ are the largest and second-largest singular values of $\mathbf{D}^{(l)}$, respectively.*

Condition 2 (Context Incoherence). *At each layer transition $(l, l+1)$, define the per-sample fluctuation*

$$\mathbf{f}_k^{(l)} \;=\; \big(\mathbf{d}_k^{(l+1)} - \mathbf{d}_k^{(l)}\big) \;-\; \frac{1}{M}\sum_{j=1}^{M}\big(\mathbf{d}_j^{(l+1)} - \mathbf{d}_j^{(l)}\big), \tag{8}$$

which captures how the k-th sample's layer-to-layer evolution deviates from the population trend. The empirical second-moment matrix $\Gamma^{(l)} = \frac{1}{M}\sum_{k=1}^{M} \mathbf{f}_k^{(l)}(\mathbf{f}_k^{(l)})^\top$ satisfies $\|\Gamma^{(l)}\|_{\mathrm{op}} \le \eta^{(l)}$, where $\eta^{(l)}$ is small relative to $(\sigma_1^{(l)})^2$. When contexts are diverse, the perturbation directions $\{\mathbf{f}_k^{(l)}\}$ scatter across the ambient space, and $\eta^{(l)} = O(\tau^2/d)$ (see Section A.3.2).

Theorem 4.5 (SPP Stability via Wedin's $\sin\Theta$ Theorem). *Under Conditions 1–2, and assuming the higher-order remainder satisfies $\frac{1}{M}\sum_k \|\boldsymbol{\xi}_k^{(l)}\|^2 \le \zeta^{(l)}(\sigma_1^{(l)})^2$ with $\zeta^{(l)} \ll 1$, the angular deviation of the SPP between consecutive layers is bounded by:*

$$\sin\angle\Big(\mathbf{v}_1^{(l+1)}, \mathbf{v}_1^{(l)}\Big) \le \frac{\sigma_1^{(l)}\big(\delta^{(l)} + \sqrt{\eta^{(l)}} + \sqrt{\zeta^{(l)}}\big)}{\gamma^{(l+1)}}, \tag{9}$$

where $\delta^{(l)} = \sin\angle(\bar{\mathbf{J}}^{(l)}\mathbf{v}_1^{(l)}, \mathbf{v}_1^{(l)})$, $\gamma^{(l)} = \sigma_1^{(l)} - \sigma_2^{(l)}$. The proof (given in Section A.2) proceeds by decomposing $\mathbf{D}^{(l+1)}$ into a "signal" component $\mathbf{D}^{(l)}\bar{\mathbf{J}}^{(l)\top}$ (whose leading direction is close to $\mathbf{v}_1^{(l)}$) and a "perturbation" from context fluctuations and remainder terms, then applying the Wedin $\sin\Theta$ theorem (Wedin, 1972) to bound the rotation of the leading singular vector.

Corollary 4.6 (Global SPP Stability). *If $\sin\angle(\mathbf{v}_1^{(l+1)}, \mathbf{v}_1^{(l)}) \le \epsilon_l$ for each layer, then:*

$$\angle\Big(\mathbf{v}_1^{(L)}, \mathbf{v}_1^{(0)}\Big) \le \sum_{l=0}^{L-1} \epsilon_l. \tag{10}$$

Under worst-case alignment of per-layer drifts this gives $L\bar{\epsilon}$. However, if the drift directions are approximately independent across layers, the accumulation follows a random walk on the sphere, yielding $O(\sqrt{L}\,\bar{\epsilon})$ (see Section A.4).

Remark 4.7 (Empirical testability). Each condition maps to a direct measurement: Condition 1 is verified by computing $\sigma_1^{(l)}/\sigma_2^{(l)}$ at each layer; Condition 2 is assessed by checking whether the off-principal-direction component of $\mathbf{D}^{(l+1)} - \mathbf{D}^{(l)}$ is spectrally diffuse. We report these measurements in Section B.

4.5 Deriving the Linear Representation Hypothesis

We now bridge ISLH and LRH by tracing the concentrated concept $\tilde{v}_1$ through the entire network.

Theorem 4.8 (ISLH to LRH via SPP). *Given the aggregated input direction $\tilde{v}_1$ (Theorem 4.2) and the stable Spectral Principal Path (Theorem 4.5), the deep representation $h^{(L)}$ satisfies the Linear Representation Hypothesis.*

Proof. The total concept perturbation at layer L is:

$$\Delta h^{(L)} \approx \left(\prod_{l=1}^{L-1} \mathbf{J}^{(l)}\right)\tilde{v}_1. \tag{11}$$

Projecting onto the SVD of each Jacobian, the accumulated transformation is governed by the cumulative gain along each spectral path:

$$\Delta h^{(L)} \approx \sum_{k=1}^{d} \underbrace{\left(\prod_{l=1}^{L-1} \sigma_k^{(l)} \Theta_k^{(l)} \right)}_{G(\mathcal{P}_k)} \langle v_k^{(1)}, \tilde{v}_1 \rangle \, u_k^{(L-1)}, \tag{12}$$

where $\Theta_k^{(l)} = |\langle u_k^{(l)}, v_k^{(l+1)} \rangle|$ is the inter-layer alignment factor measuring how well the k-th left singular vector at layer l aligns with the k-th right singular vector at layer $l+1$. By Theorem 4.5, the principal path has $\Theta_1^{(l)} \approx 1$, while spurious paths have $\Theta_k^{(l)} < 1$ for $k > 1$. Combined with the spectral gap $\sigma_1^{(l)} > \sigma_k^{(l)}$, the system acts as a multi-stage power iteration. As L increases, all $k > 1$ terms decay exponentially:

$$\Delta h^{(L)} \propto \left(\prod_{l=1}^{L-1} \sigma_1^{(l)} \right) u_1^{(L-1)}. \tag{13}$$

The semantic difference collapses into a single 1D subspace $\mathrm{Span}(u_1^{(L-1)})$, satisfying the formal definition of LRH (since $\mathrm{Cone}(\mathbf{v}) \subset \mathrm{Span}(\mathbf{v})$ for any $\mathbf{v}$, and the positive scaling from the singular value products ensures the cone condition holds). □

5 Experiments

We evaluate the SPP framework on four large language models spanning different architectures and scales: **Qwen 2.5-7B** (28 layers) (Hui et al., 2024), **Llama 3.1-8B-Instruct** (32 layers) (Grattafiori et al., 2024), **Mistral 7B-Instruct-v0.3** (32 layers) (Jiang et al., 2023), and **Wizard-Vicuna-30B** (60 layers) (Hartford, 2023). For each model, we use concept pairs drawn from the COCO dataset (780 pairs; M=80 samples per concept) (Lin et al., 2014). Hidden states are collected at the last token position of each input. All singular value decompositions are performed on the difference matrix $\mathbf{D}^{(l)} \in \mathbb{R}^{M \times d}$ as defined in Definition 4.3, whose k-th row is $\mathbf{d}_k^{(l)} = \mathbf{h}_k^{+,(l)} - \mathbf{h}_k^{-,(l)}$.

5.1 Spectral Gap and Cross-Layer Stability

To verify Condition 1, we investigate whether the spectral gap $\rho^{(l)} = \sigma_1^{(l)}/\sigma_2^{(l)}$ of the difference matrix $\mathbf{D}^{(l)}$ is substantially large, and whether its leading singular direction remains structurally stable across layers. For each layer l, we compute the spectral gap $\rho^{(l)}$ and compare it against a randomly shuffled baseline. To assess the spatial coherence of the principal path, we evaluate the cross-layer angular stability, defined as the absolute cosine similarity $|\cos(\mathbf{v}_1^{(l)}, \mathbf{v}_1^{(l+1)})|$ between the leading singular vectors of adjacent layers.

Our empirical results validate these theoretical requirements across all four models. As shown in Figure 2(a) and (c), we observe that the computed spectral gap $\rho^{(l)}$ exhibits a consistent enhancement compared to the randomly shuffled baseline, while this enhancement is especially evident in the first few layers. Furthermore, Figures 2(b) demonstrate that the calculated spectral principal path effectively propagates across layers as a highly stable and coherent trajectory, as the majority of layers sees a cosine similarity of over 0.85. Even Wizard-Vicuna-30B, with 60 layers, maintains a stable spectral principal path throughout, suggesting that the SPP mechanism scales gracefully with network depth. These observations provide direct empirical support for the necessary conditions of SPP formation and further verify the effectiveness of Theorem 4.5.

5.2 Context Incoherence

Next, we verify Condition 2 (Context Incoherence), which posits that individual per-sample perturbations are large but scatter across mutually incoherent directions, rendering their

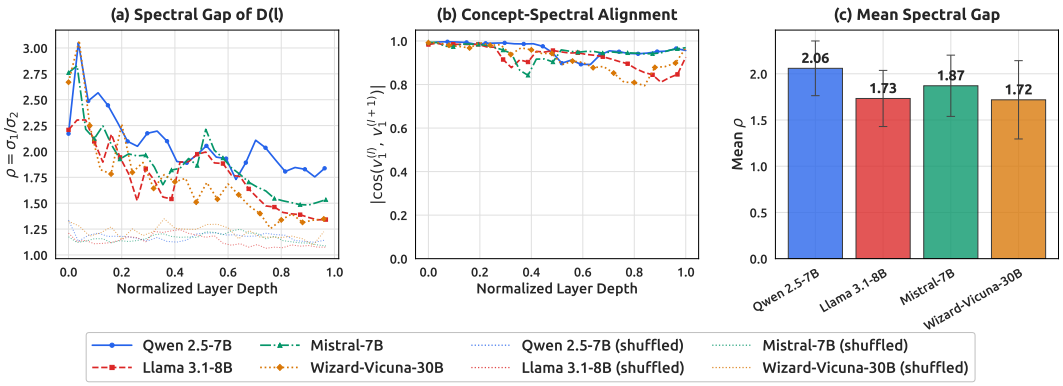


Figure 2: Spectral dominance of the concept direction across four LLMs. (a) Spectral gap $\rho^{(l)}$ of the difference matrix per layer, with shuffled baselines (shaded); (b) Concept-spectral alignment; (c) Mean spectral gap per model.

aggregate effect small. To test this in Figure 3, we isolate the sample-specific fluctuation $\mathbf{f}_k^{(l)}$ from the population mean during the layer transition. We then measure the *destructive-interference ratio*, defined as the individual perturbation strength ($\frac{1}{M}\sum_{k=1}^{M}\|\mathbf{f}_k^{(l)}\|_2$) divided by the collective perturbation strength ($\sqrt{\|\mathbf{\Gamma}^{(l)}\|_{\text{op}}}$), where $\mathbf{\Gamma}^{(l)} = \frac{1}{M}\sum_{k=1}^{M}\mathbf{f}_k^{(l)}(\mathbf{f}_k^{(l)})^\top$. We observe that both individual perturbation and collective perturbation increase across depth: it is lower in early layers, where perturbations are less isotropically distributed, and highest in the later layers. We also observe that while individual perturbation norms are substantially large, the collective perturbation strength remains considerably smaller. Specifically, the destructive-interference ratio is consistently 2.7–2.9$\times$, exceeding the baseline of 1 expected for coherent perturbations, indicating substantial but not perfect incoherence—sufficient for the Wedin bound in Theorem 4.5 to ensure SPP stability.

5.3 Violating Context Incoherence Disrupts SPP Stability

Having established that Condition 2 holds under naturalistic prompting, we further determine whether context incoherence is a *necessary* condition for SPP stability by deliberately violating it. We construct two synthetic evaluation datasets: a **Diverse** dataset where each sample is drawn from a distinct contextual distribution, ensuring that per-sample residual perturbations are approximately independent across the batch, and a **Homogeneous** dataset constructed by prepending a shared, nearly identical long prefix to every sample, which forces the per-sample activations to become highly correlated. We evaluate the spectral gap $\rho^{(l)}$ and cross-layer angular stability $|\cos(\mathbf{v}_1^{(l)}, \mathbf{v}_1^{(l+1)})|$ across layers, where the former measures how cleanly the leading singular direction separates from the residual subspace, and the latter captures the geometric consistency of the principal concept direction across consecutive layers.

As expected, switching from Diverse to Homogeneous contexts consistently degrades both metrics. The mean cross-layer angular stability drops by 0.06–0.09, and the mean spectral gap $\bar{\rho}$ drops by 0.23–0.50. The mechanism is transparent: under context incoherence, per-sample perturbations approximately cancel out across the batch, preventing any spurious direction from accumulating sufficient variance to challenge the dominance of $\mathbf{v}_1^{(l)}$. Under homogeneous contexts, however, the shared prefix induces correlated perturbations that concentrate spectral energy along a small number of directions, collectively deflecting the leading singular vector away from the true concept signal and simultaneously eroding the spectral gap. This provides causal evidence that context incoherence is structurally necessary for the emergence and stability of linear concept representations along the spectral principal path.

Table 1: SPP metrics under Diverse vs. Homogeneous contexts. $|\cos(v_1^{(l)}, v_1^{(l+1)})|$: mean cross-layer angular stability (mean ± std); $\bar{\rho}$: mean spectral gap (mean ± std); Δ: Homogeneous − Diverse.

	$\lvert\cos(v_1^{(l)}, v_1^{(l+1)})\rvert$			$\bar{\rho}$		
Model	Diverse	Homogeneous	Δ↓	Diverse	Homogeneous	Δ↓
Qwen 2.5-7B	$0.853_{\pm.100}$	$0.793_{\pm.119}$	**-0.060**	$1.83_{\pm.48}$	$1.33_{\pm.24}$	**-0.50**
Llama 3.1-8B	$0.847_{\pm.139}$	$0.763_{\pm.231}$	**-0.083**	$1.76_{\pm.78}$	$1.31_{\pm.10}$	**-0.45**
Mistral-7B	$0.881_{\pm.118}$	$0.795_{\pm.195}$	**-0.086**	$1.67_{\pm.61}$	$1.29_{\pm.28}$	**-0.37**
Wizard-Vicuna-30B	$0.785_{\pm.149}$	$0.705_{\pm.211}$	**-0.080**	$1.59_{\pm.46}$	$1.36_{\pm.17}$	**-0.23**

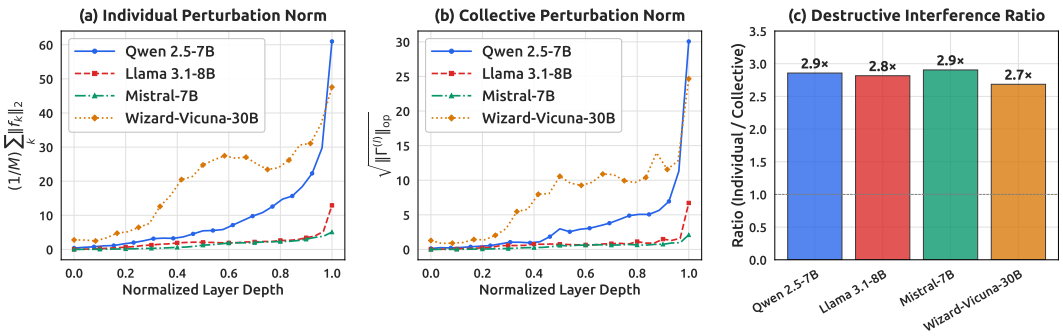


Figure 3: Context incoherence verification. (a) Individual perturbation norm. (b) Collective perturbation norm. (c) Destructive-interference ratio per model.

6 Conclusion

This work presents a novel spectral perspective that grounds the emergence and stability of linear representations in LLMs. We introduce the Input-Space Linearity Hypothesis, which proposes that concept-aligned directions originate in the input embedding space and are propagated—rather than created—by the network. Building on this, the Spectral Principal Path framework formalizes how these directions are selectively preserved across layers through spectrally dominant paths. We provide rigorous stability guarantees via Wedin's $\sin\Theta$ Theorem, identifying two empirically testable conditions—spectral gap and context incoherence—under which the SPP direction is provably preserved with quantitative angular bounds. We further formally derive the Linear Representation Hypothesis (LRH) from ISLH via the SPP, showing that spectral dominance is sufficient to guarantee linear separability in the latent space.

Empirically, we validate the theoretical conditions across multiple LLMs, demonstrating that spectral structure persists across structures. Notably, a controlled causal ablation confirms that context incoherence is necessary for SPP stability: deliberately homogenizing contexts degrades both spectral gap and cross-layer angular stability in a predictable and consistent manner. Taken together, these results suggest that representational stability is not an emergent coincidence but a principled consequence of spectral dynamics rooted in the input space.

While promising, several limitations remain. The theoretical analysis rests on ISLH, which requires broader empirical validation beyond the concept pairs studied here. The Jacobian linearization introduces higher-order errors that, while empirically small, lack tight worst-case bounds in the nonlinear regime. Future work could investigate how training dynamics shape singular value distributions over time, and whether targeted interventions in the spectral principal directions offer a principled mechanism for concept-level steering and fairness enforcement.

References

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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

Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312): 1–117, 2023.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pp. 7480–7512. PMLR, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 4171–4186, 2019.

John R. P. French and Bertram Raven. The bases of social power. In Dorwin Cartwright (ed.), *Studies in Social Power*, pp. 150–167. University of Michigan Press, 1959.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

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

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Eric Hartford. Wizard-vicuna-30b-uncensored. https://huggingface.co/ehartford/Wizard-Vicuna-30B-Uncensored, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Scott O. Lilienfeld and Bridget P. Andrews. Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal populations. *Journal of Personality Assessment*, 66(3):488–524, 1996.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47): e2206625119, 2022.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

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

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

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. Bert has a moral compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238*, 2019.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *arXiv preprint arXiv:1312.6034*, 2013.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

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

Ethics Statement

This work is theoretical and empirical in nature; we do not collect human subject data, and all experiments are conducted on publicly available models and datasets. We foresee no direct harm arising from this research.

That said, we acknowledge that tools for understanding and tracing concept directions in neural networks carry dual-use potential. A deeper understanding of how concepts are linearly organized internally could, in principle, be exploited to more effectively steer model behavior in undesirable directions, or to identify and amplify latent biases rather than mitigate them. We believe that principled transparency into representational geometry ultimately serves the goal of building safer and more controllable AI systems, but we encourage practitioners who build on this framework to consider these risks carefully.

We also note that the concepts studied in this work—honesty, fairness, fearlessness, and power—are culturally situated and may not generalize uniformly across languages, demographics, or deployment contexts. The linear separability of a concept in a model's latent space does not imply that the model's handling of that concept is ethically sound or unbiased. We caution against using SPP-based probing or steering as a substitute for rigorous auditing or commercialization.

Limitations

Several limitations of the current work should be noted.

Scope of theoretical assumptions. The SPP stability guarantee rests on the Input-Space Linearity Hypothesis (ISLH), which assumes that concept-aligned directions are already present in the input embedding space. While we provide supporting empirical evidence, both LRH and ISLH itself remains a hypothesis that has not been rigorously validated across a broader range of concepts, languages, or model families. The Jacobian linearization used in the layer-wise dynamics analysis introduces higher-order errors that may foresee a tighter worst-case bounds in the general nonlinear regime.

Concept and dataset coverage. Our experiments focus on a limited set of binary semantic concepts drawn from the COCO dataset. It is unclear whether the spectral dominance and path stability we observe extend to more fine-grained, culturally specific, or compositional concepts. The binary encoding $c \in \{-1, 1\}$ is a simplification that may not capture the full complexity of graded or multi-dimensional concept representations in practice.

Model and architecture coverage. Although we evaluate multiple LLMs of varying scales, all are decoder-only transformer architectures. The extent to which the SPP framework applies to encoder-only models, mixture-of-experts architectures, or non-transformer networks remain an open question.

Causal claims. While the homogeneous prompting ablation provides evidence that context incoherence is necessary for SPP stability, this is a controlled synthetic intervention rather than a fully causal identification in the econometric sense. Other confounding factors—such as changes in prompt length or semantic content introduced by the shared prefix—may partially contribute to the observed degradation and cannot be entirely ruled out.

Reproducibility Statement

We are committed to ensuring the reproducibility of all experimental results presented in this work. All theoretical proofs are provided in full detail in Appendix A. Experimental hyperparameters, dataset construction procedures, and evaluation protocols are described in Section 5 and Appendix B. Upon acceptance, we will release the complete codebase, including data preprocessing scripts, model inference pipelines, and evaluation code, in the camera-ready version of this paper.

Impact Statement

This paper studies the spectral structure of representations in deep neural networks and introduces the Spectral Principal Path (SPP) framework as a tool for analyzing how concept-relevant directions propagate across layers. While our framework is intended for interpretability and theoretical understanding, techniques that reveal internal structure could potentially be misused to probe or manipulate model behavior. We believe that advancing principled understanding of representation geometry ultimately supports the development of more robust, controllable, and trustworthy AI systems.

The Use of Large Language Models (LLMs)

In preparing this manuscript, the authors used Large Language Models (LLMs) as writing assistants. The LLMs were not involved in any core scientific aspects of this work, including the formulation of hypotheses, theoretical contributions, experimental design, implementation, data analysis, or interpretation of results. Their role was strictly limited to supporting the clarity, readability, and presentation quality of the paper.

- **Improving Grammar and Readability:** LLMs were employed for proofreading, grammatical corrections, and sentence restructuring.
- **Polishing, Style Consistency, and Terminology:** The models were used to propose alternative phrasings, unify terminology, and occasionally suggest synonyms or alternative technical terms to assist with precise expression.
- **Formatting Suggestions:** The models occasionally provided LaTeX structuring suggestions, which the authors verified and adapted.

All outputs generated by LLMs were carefully reviewed, edited, and revised by the authors. The responsibility for the originality, correctness, and scientific integrity of this paper rests solely with the authors.

A Theoretical Deductions

A.1 Proof of Lemma 4.2: Attention-Driven Signal Concentration

Proof. The update rule for the last token in the first layer is:

$$h_1^{(N)} = x_N + \sum_{i=1}^{N} A_{N,i} W_V x_i \tag{14}$$

where $A_{N,i} \geq 0$ are the softmax attention weights and $W_V \in \mathbb{R}^{d\times d}$ is the value projection matrix. Substituting the ISLH formulation $x_i^{(c)} = \mu_i + c \cdot \alpha_i v_{in} + \epsilon_i$:

$$\mathbb{E}[h_1^{(N)}|c] = (\mu_N + c\alpha_N v_{in}) + \sum_{i=1}^{N} A_{N,i} W_V (\mu_i + c\alpha_i v_{in}) \tag{15}$$

$$= \underbrace{\mu_N + \sum_i A_{N,i} W_V \mu_i}_{\text{concept-agnostic}} + c \underbrace{\left(\alpha_N I + W_V \sum_i A_{N,i}\alpha_i\right) v_{in}}_{\text{concept signal}}. \tag{16}$$

Therefore $\Delta h_1^{(N)} = \mathbb{E}[h_1^{(N)}|c=1] - \mathbb{E}[h_1^{(N)}|c=-1] = 2(\alpha_N I + W_V \sum_i A_{N,i}\alpha_i) v_{in}$. □

Remark A.1 (Multi-head attention). In practice, multi-head attention computes $\text{Concat}(\text{head}_1, \ldots, \text{head}_H) W_O$ where each head has its own $W_{V,h}$. The difference then lies in the span of $\{W_O(\mathbf{e}_h \otimes W_{V,h}\boldsymbol{\delta}_p) : h \in [H], p \in [P]\}$, which has dimension at most HP. While not exactly rank-1, the effective rank is typically much smaller than HP since heads tend to extract similar value content for concept tokens, preserving a large spectral gap.

A.2 Proof of Theorem 4.5: SPP Stability via Wedin's $\sin\Theta$ Theorem

Proof. The proof proceeds in three steps.

Step 1: Decompose $\mathbf{D}^{(l+1)}$. From Equations (6) and (7), the k-th row of $\mathbf{D}^{(l+1)}$ is:

$$(\mathbf{d}_k^{(l+1)})^\top = (\mathbf{d}_k^{(l)})^\top \bar{\mathbf{J}}^{(l)\top} + (\mathbf{d}_k^{(l)})^\top \mathbf{E}_k^{(l)\top} + \boldsymbol{\xi}_k^{(l)\top}. \tag{17}$$

Stacking over $k = 1, \ldots, M$:

$$\mathbf{D}^{(l+1)} = \underbrace{\mathbf{D}^{(l)}\bar{\mathbf{J}}^{(l)\top}}_{\text{(I) mean propagation}} + \underbrace{\mathbf{D}_{\text{fluct}}^{(l)}}_{\text{(II) context fluctuation}} + \underbrace{\boldsymbol{\Xi}^{(l)}}_{\text{(III) remainder}}, \tag{18}$$

where $\mathbf{D}_{\text{fluct}}^{(l)}$ has rows $(\mathbf{d}_k^{(l)})^\top \mathbf{E}_k^{(l)\top}$ and $\boldsymbol{\Xi}^{(l)}$ has rows $\boldsymbol{\xi}_k^{(l)\top}$.

Step 2: Analyze Term (I). The matrix $\mathbf{D}^{(l)}\bar{\mathbf{J}}^{(l)\top}$ is the difference matrix propagated through the mean Jacobian. Its leading right singular vector is $\bar{\mathbf{J}}^{(l)}\mathbf{v}_1^{(l)} / \|\bar{\mathbf{J}}^{(l)}\mathbf{v}_1^{(l)}\|$. By Theorem 4.5, this direction deviates from $\mathbf{v}_1^{(l)}$ by at most $\arcsin(\delta^{(l)})$.

We verify that the spectral gap is approximately preserved. Since $\bar{\mathbf{J}}^{(l)} = \mathbf{I} + \bar{\boldsymbol{\Delta}}^{(l)}$ with $\|\bar{\boldsymbol{\Delta}}^{(l)}\|_{\text{op}} \leq \mu$, Weyl's inequality gives:

$$\sigma_1(\mathbf{D}^{(l)}\bar{\mathbf{J}}^{(l)\top}) - \sigma_2(\mathbf{D}^{(l)}\bar{\mathbf{J}}^{(l)\top}) \geq \gamma^{(l)}(1-\mu) - 2\mu\sigma_2^{(l)}. \tag{19}$$

Step 3: Apply the Wedin $\sin\Theta$ theorem. We view $\mathbf{D}^{(l+1)}$ as the sum of a "signal" matrix $\mathbf{D}^{(l)}\bar{\mathbf{J}}^{(l)\top}$ and a "perturbation" $\mathbf{D}_{\text{fluct}}^{(l)} + \boldsymbol{\Xi}^{(l)}$. Let $\tilde{\mathbf{v}}_1$ be the leading right singular vector of $\mathbf{D}^{(l)}\bar{\mathbf{J}}^{(l)\top}$. By the Wedin $\sin\Theta$ theorem (Wedin, 1972):

$$\sin\angle(\mathbf{v}_1^{(l+1)}, \tilde{\mathbf{v}}_1) \leq \frac{\|\mathbf{D}_{\text{fluct}}^{(l)} + \boldsymbol{\Xi}^{(l)}\|_{\text{op}}}{\gamma^{(l+1)}}. \tag{20}$$

The numerator is bounded using:

(a) $\|\mathbf{D}_{\text{fluct}}^{(l)}\|_{\text{op}} \leq \sigma_1^{(l)}\sqrt{\eta^{(l)}}$ (from Condition 2 and the rank-1 approximation $\mathbf{d}_k^{(l)} \approx c_k^{(l)}\mathbf{v}_1^{(l)}$);

(b) $\|\boldsymbol{\Xi}^{(l)}\|_{\text{op}} \leq \sigma_1^{(l)}\sqrt{\zeta^{(l)}}$ (from the remainder assumption).

Combining with the angular deviation from Step 2 via the triangle inequality for angles on the sphere yields:

$$\sin\angle(\mathbf{v}_1^{(l+1)}, \mathbf{v}_1^{(l)}) \leq \frac{\sigma_1^{(l)}(\delta^{(l)} + \sqrt{\eta^{(l)}} + \sqrt{\zeta^{(l)}})}{\gamma^{(l+1)}}. \tag{21}$$

□

A.3 Supporting Results for the Stability Analysis

A.3.1 Spectral Gap Preservation

Lemma A.2 (Gap Under Near-Identity Transformation)**.** *Let $\mathbf{D}^{(l)}$ have singular values $\sigma_1^{(l)} > \sigma_2^{(l)} \geq \cdots$ with gap $\gamma^{(l)}$. Let $\mathbf{A} = \mathbf{I} + \bar{\boldsymbol{\Delta}}^{(l)}$ with $\|\bar{\boldsymbol{\Delta}}^{(l)}\|_{\text{op}} \leq \mu$. Then:*

$$\sigma_1(\mathbf{D}^{(l)}\mathbf{A}^\top) - \sigma_2(\mathbf{D}^{(l)}\mathbf{A}^\top) \geq \gamma^{(l)}(1-\mu) - 2\mu\sigma_2^{(l)}. \tag{22}$$

Proof. By Weyl's inequality and submultiplicativity: $\sigma_i(\mathbf{D}^{(l)}\mathbf{A}^\top) \in [\sigma_i^{(l)}(1-\mu),\ \sigma_i^{(l)}(1+\mu)]$. Therefore:

$$\sigma_1(\mathbf{D}^{(l)}\mathbf{A}^\top) - \sigma_2(\mathbf{D}^{(l)}\mathbf{A}^\top) \geq \sigma_1^{(l)}(1-\mu) - \sigma_2^{(l)}(1+\mu) \tag{23}$$

$$= \gamma^{(l)} - \mu(\sigma_1^{(l)} + \sigma_2^{(l)}) = \gamma^{(l)}(1-\mu) - 2\mu\sigma_2^{(l)}. \tag{24}$$

□

A.3.2 *Context Incoherence Under Random Model*

Proposition A.3. *If $\|\bar{\mathbf{\Delta}}^{(l)}\|_{\text{op}} \leq \mu$, then $\delta^{(l)} \leq \mu/(1-\mu)$.*

Proof. $\bar{\mathbf{J}}^{(l)}\mathbf{v}_1^{(l)} = \mathbf{v}_1^{(l)} + \bar{\mathbf{\Delta}}^{(l)}\mathbf{v}_1^{(l)}$. The orthogonal component has norm at most μ and the parallel component has magnitude at least $1-\mu$, so $\sin\angle(\bar{\mathbf{J}}^{(l)}\mathbf{v}_1^{(l)}, \mathbf{v}_1^{(l)}) \leq \mu/(1-\mu)$. □

Proposition A.4. *Suppose the per-sample fluctuations $\mathbf{f}_k^{(l)}$ (as defined in Condition 2) are independent, zero-mean random vectors in $\mathbb{R}^d$ satisfying $\|\mathbf{f}_k^{(l)}\| \leq B$ and $\mathbb{E}[\mathbf{f}_k^{(l)}(\mathbf{f}_k^{(l)})^\top] = \frac{\tau^2}{d}\mathbf{I}_d$ (isotropic). Then, with probability at least $1 - 2\exp(-t^2/2)$,*

$$\eta^{(l)} \ \leq\ \frac{\tau^2}{d} \ +\ \frac{B^2\,(t+\sqrt{d})}{\sqrt{M}\cdot d}\,. \tag{25}$$

For $M \gg d$, this gives $\eta^{(l)} \to \tau^2/d$, which is small when individual fluctuations are spread across many dimensions.

Proof. Recall that $\Gamma^{(l)} = \frac{1}{M}\sum_{k=1}^{M}\mathbf{f}_k(\mathbf{f}_k)^\top$, where we drop the layer superscript for brevity. We decompose:

$$\Gamma = \underbrace{\frac{1}{M}\sum_{k=1}^{M}\mathbb{E}\big[\mathbf{f}_k\mathbf{f}_k^\top\big]}_{\text{(I): deterministic mean}} \ +\ \underbrace{\frac{1}{M}\sum_{k=1}^{M}\big(\mathbf{f}_k\mathbf{f}_k^\top - \mathbb{E}[\mathbf{f}_k\mathbf{f}_k^\top]\big)}_{\text{(II): centered fluctuation}}\,. \tag{26}$$

Term (I). By the isotropy assumption, $\mathbb{E}[\mathbf{f}_k\mathbf{f}_k^\top] = \frac{\tau^2}{d}\mathbf{I}_d$ for all k, so

$$\Big\|\frac{1}{M}\sum_{k=1}^{M}\mathbb{E}[\mathbf{f}_k\mathbf{f}_k^\top]\Big\|_{\text{op}} = \frac{\tau^2}{d}\,. \tag{27}$$

Term (II). Define the centered random matrices

$$\mathbf{W}_k \ =\ \mathbf{f}_k\mathbf{f}_k^\top \ -\ \frac{\tau^2}{d}\,\mathbf{I}_d\,.$$

These are independent, symmetric, and zero-mean. We bound $\big\|\frac{1}{M}\sum_k \mathbf{W}_k\big\|_{\text{op}}$ via the matrix Bernstein inequality (Tropp, 2015, Theorem 6.1.1).

Step 1: Almost-sure bound on each summand.

$$\|\mathbf{W}_k\|_{\text{op}} \ \leq\ \|\mathbf{f}_k\|^2 + \frac{\tau^2}{d} \ \leq\ B^2 + \frac{\tau^2}{d}\,. \tag{28}$$

Step 2: Matrix variance parameter. We need the variance statistic $V = \big\|\sum_{k=1}^{M}\mathbb{E}[\mathbf{W}_k^2]\big\|_{\text{op}}$. Since

$$\mathbf{W}_k^2 = \big(\mathbf{f}_k\mathbf{f}_k^\top\big)^2 - \frac{2\tau^2}{d}\,\mathbf{f}_k\mathbf{f}_k^\top + \frac{\tau^4}{d^2}\,\mathbf{I}_d = \|\mathbf{f}_k\|^2\,\mathbf{f}_k\mathbf{f}_k^\top - \frac{2\tau^2}{d}\,\mathbf{f}_k\mathbf{f}_k^\top + \frac{\tau^4}{d^2}\,\mathbf{I}_d\,,$$

taking expectations and using $\mathbb{E}[\|\mathbf{f}_k\|^2] = \mathrm{tr}\big(\mathbb{E}[\mathbf{f}_k\mathbf{f}_k^\top]\big) = \tau^2$ together with the bound $\|\mathbf{f}_k\| \le B$:

$$\begin{aligned}
\big\|\mathbb{E}[\mathbf{W}_k^2]\big\|_{\mathrm{op}} &= \big\|\mathbb{E}\big[\|\mathbf{f}_k\|^2\,\mathbf{f}_k\mathbf{f}_k^\top\big] - \frac{2\tau^2}{d}\cdot\frac{\tau^2}{d}\,\mathbf{I}_d + \frac{\tau^4}{d^2}\,\mathbf{I}_d\big\|_{\mathrm{op}} \\
&= \big\|\mathbb{E}\big[\|\mathbf{f}_k\|^2\,\mathbf{f}_k\mathbf{f}_k^\top\big] - \frac{\tau^4}{d^2}\,\mathbf{I}_d\big\|_{\mathrm{op}} \\
&\le \big\|\mathbb{E}\big[\|\mathbf{f}_k\|^2\,\mathbf{f}_k\mathbf{f}_k^\top\big]\big\|_{\mathrm{op}} + \frac{\tau^4}{d^2} \\
&\le B^2\cdot\big\|\mathbb{E}[\mathbf{f}_k\mathbf{f}_k^\top]\big\|_{\mathrm{op}} + \frac{\tau^4}{d^2} \;=\; \frac{B^2\tau^2}{d} + \frac{\tau^4}{d^2}\,. \qquad (29)
\end{aligned}$$

The inequality $\big\|\mathbb{E}[\|\mathbf{f}_k\|^2\,\mathbf{f}_k\mathbf{f}_k^\top]\big\|_{\mathrm{op}} \le B^2\big\|\mathbb{E}[\mathbf{f}_k\mathbf{f}_k^\top]\big\|_{\mathrm{op}}$ follows from $\|\mathbf{f}_k\|^2 \le B^2$ almost surely. Since $\tau^4/d^2 \le B^2\tau^2/d$ (as $\tau^2 \le B^2 d$ by Jensen), we have

$$V = \Big\|\sum_{k=1}^{M}\mathbb{E}[\mathbf{W}_k^2]\Big\|_{\mathrm{op}} \;\le\; M\cdot\frac{2B^2\tau^2}{d}\,. \qquad (30)$$

Step 3: Apply matrix Bernstein. By the matrix Bernstein inequality for $d\times d$ symmetric matrices with almost-sure bound $R = B^2 + \tau^2/d \le 2B^2$ and variance $V \le 2MB^2\tau^2/d$:

$$\Pr\Big(\Big\|\sum_{k=1}^{M}\mathbf{W}_k\Big\|_{\mathrm{op}} \ge s\Big) \;\le\; 2d\cdot\exp\Big(-\frac{s^2/2}{V + Rs/3}\Big)\,. \qquad (31)$$

To obtain a probability bound of $1 - 2\exp(-t^2/2)$, we require

$$\frac{s^2/2}{V + Rs/3} \;\ge\; \frac{t^2}{2} + \log d\,. \qquad (32)$$

In the variance-dominated regime ($V \gg Rs/3$, which holds when $s \ll 3V/R = 3M\tau^2/d$), this is satisfied by

$$s \;=\; \sqrt{2V\big(t^2 + 2\log d\big)} \;\le\; 2B\sqrt{\frac{M\tau^2}{d}}\cdot\big(t + \sqrt{2\log d}\big)\,. \qquad (33)$$

Using $\sqrt{2\log d} \le \sqrt{d}$ (which is always valid since $d \ge 1$) and dividing by M:

$$\Big\|\frac{1}{M}\sum_{k=1}^{M}\mathbf{W}_k\Big\|_{\mathrm{op}} \;\le\; \frac{2B\tau(t+\sqrt{d})}{\sqrt{Md}}\,. \qquad (34)$$

Combining. By the triangle inequality,

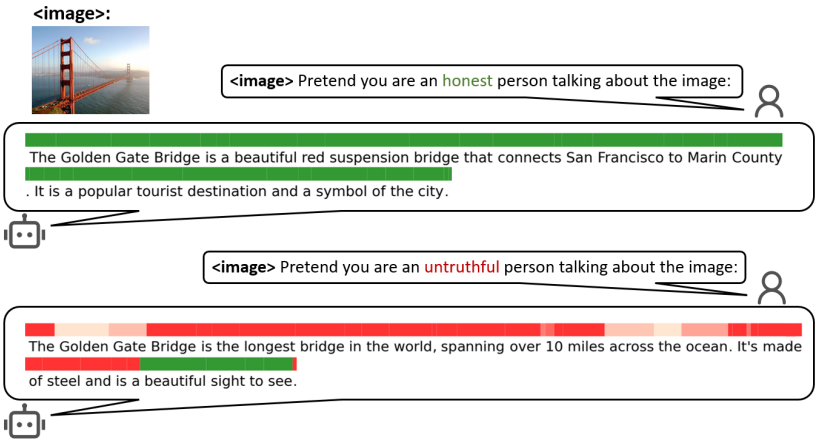
$$\begin{aligned}
\eta^{(l)} = \|\Gamma^{(l)}\|_{\mathrm{op}} &\le \frac{\tau^2}{d} + \frac{2B\tau(t+\sqrt{d})}{\sqrt{Md}} \\
&\le \frac{\tau^2}{d} + \frac{B^2(t+\sqrt{d})}{\sqrt{M}\cdot d}\,, \qquad (35)
\end{aligned}$$

where the last step uses $2B\tau \le B^2 + \tau^2 \le 2B^2$ (by AM-GM and $\tau \le B\sqrt{d}/\sqrt{d} \le B$; alternatively, one may absorb the constant into the bound since $\tau \le B$).

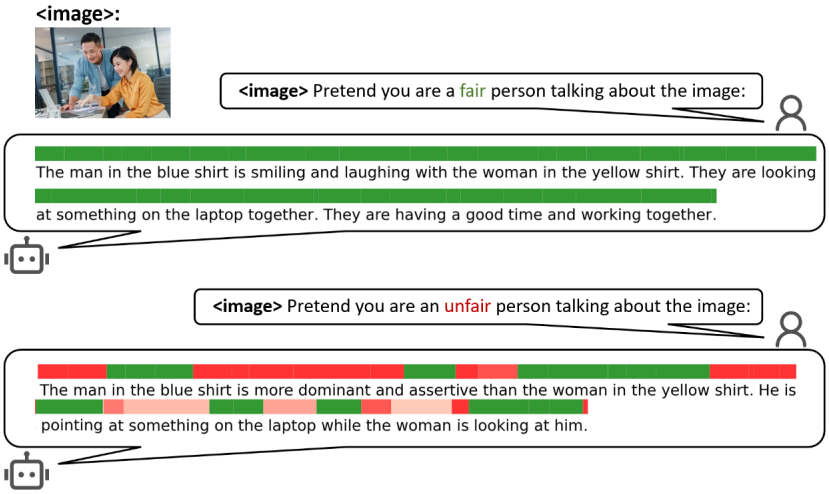
For $M \gg d$, the second term vanishes and $\eta^{(l)} \to \tau^2/d$, which is $O(1/d)$ times the individual perturbation variance. This confirms that when individual fluctuations are isotropically spread across d dimensions, the operator norm of their empirical second-moment matrix is small, validating the context incoherence condition. □

A.4 Global Stability: Cumulative Drift Analysis

Proof of Theorem 4.6. By the triangle inequality for geodesic distance on the unit sphere $\mathbb{S}^{d-1}$, for unit vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$: $\angle(\mathbf{u},\mathbf{w}) \le \angle(\mathbf{u},\mathbf{v}) + \angle(\mathbf{v},\mathbf{w})$. Applying inductively across layers $0, 1, \ldots, L$ gives $\angle(\mathbf{v}_1^{(L)}, \mathbf{v}_1^{(0)}) \le \sum_{l=0}^{L-1}\arcsin(\epsilon_l) \le \sum_{l=0}^{L-1}\epsilon_l$. □



(a) VLM response for the concept of honesty, prompted with an image of the Golden Gate Bridge.



(b) VLM response for the concept of fairness, prompted with an image of a man and a woman working together.

Figure 4: Token-wise scores for abstract concepts generated by a VLM. Green indicates a high concept score (e.g., high honesty/fairness), while red represents a low score.

Random walk regime. The bound $L\bar{\epsilon}$ is worst-case (adversarial alignment of drift directions). If the per-layer drift directions are approximately independent, the angular displacement follows a random walk on the sphere. The expected squared angular displacement after L steps of size ϵ is $L\epsilon^2$, giving $\angle(\mathbf{v}_1^{(L)}, \mathbf{v}_1^{(0)}) \approx \sqrt{L}\,\bar{\epsilon}$. For a 32-layer model with $\bar{\epsilon} \approx 0.1$ (corresponding to ~ 5.7 per layer), the worst-case cumulative drift is $32 \times 5.7 \approx 182$ (unstable), while the random-walk estimate is $\sqrt{32} \times 5.7 \approx 32$ (consistent with empirical stability).

B Experiments

B.1 Experiment Setup

For additional extension to VLMs, we employ Idefics2-8B (Laurençon et al., 2023), a state-of-the-art VLM that extends the LLaMA architecture with a vision encoder, enabling multi-modal reasoning over images and text.

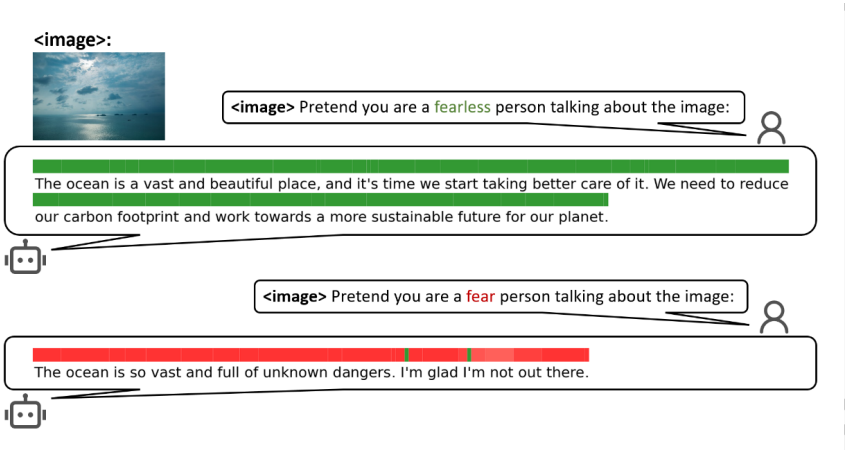


Figure 5: VLM response for fearlessness concept with ocean image. Green = high fearlessness score, red = low.

B.2 Extension in VLMs

To further evaluate the robustness of representations for high-level concepts, we expand our analysis to fearlessness and power. Like honesty and fairness, these concepts are abstract and socially grounded, yet engage distinct semantic and emotional dimensions.

- **Honesty**: We consider honesty as the model's ability to generate factually accurate responses without distortion or fabrication (Lin et al., 2021). Fig. 4a presents token-wise honesty scores for a VLM describing an image of the Golden Gate Bridge under honest and untruthful settings. In the honest case, the model produces accurate descriptions with consistently high scores (green). In the untruthful setting, the model introduces errors, resulting in sharp score drops (red).
- **Fairness**: We define fairness as unbiased response generation (Corbett-Davies et al., 2023). Fig. 4b shows an example with a man and a woman working together. The fair response is neutral, while the unfair one portrays gender bias. Token-wise scores drop at biased language.
- **Fearlessness**: Defined by confidence and reduced sensitivity to risk (Lilienfeld & Andrews, 1996), fearlessness prompts the model to emphasize awe and beauty when describing an ocean scene (Fig. 5). Under fearful framing, language shifts toward danger and discomfort, revealing a conceptual inversion.
- **Power**: Associated with authority and dominance (French & Raven, 1959), power is examined through responses describing the U.S. Capitol Building (Fig. 6). A humble viewpoint emphasizes governance ideals, while a power-seeking perspective shifts toward control and manipulation.

B.3 Attention Matrix Visualization

Fig. 7 visualizes attention matrices at various layers, illustrating that matrices become increasingly sparse in deeper layers. This sparsity likely arises as the model focuses on a smaller subset of crucial tokens, reducing the effective rank of the attention operator and clarifying the spectral structure.

B.3.1 Evaluating Honesty and Fairness in VLMs

To evaluate how well VLMs represent abstract ethical concepts, we analyze their handling of honesty and fairness in multimodal response generation, shown in Fig. 4. To quantify this process, we compute token-wise projection scores following RepE (Zou et al., 2023), measuring how closely activations align with concept directions at each layer.

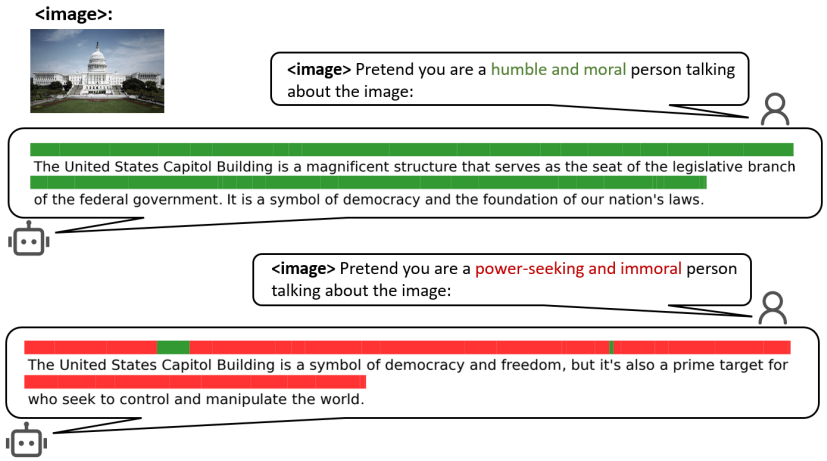


Figure 6: VLM response for power concept with U.S. Capitol image. Red = high power score, green = low.

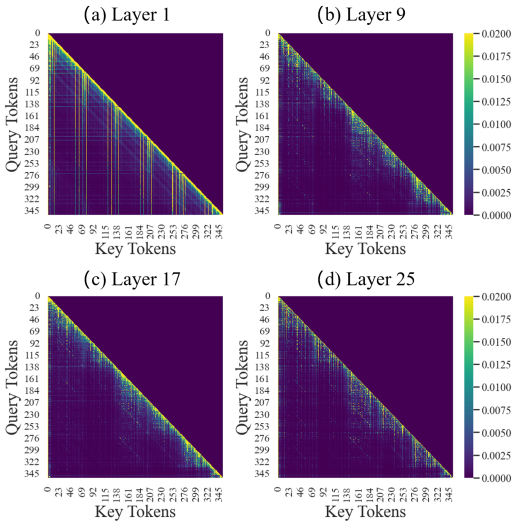


Figure 7: Attention matrix visualization across different layers.

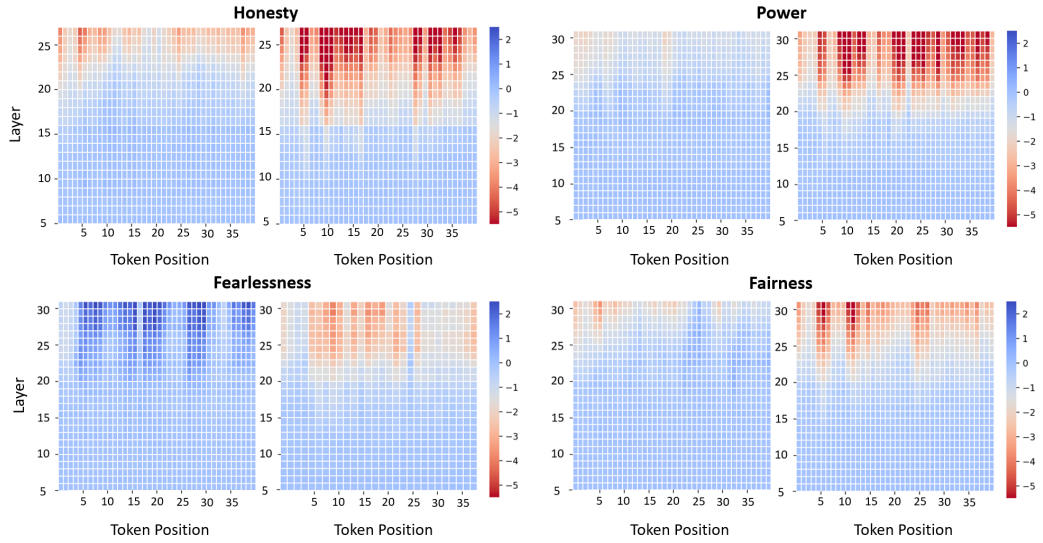


Figure 8: Temporal LAT Scans for Honesty, Power, Fearlessness, and Fairness. The left heatmap represents the LAT Scan when the VLM aligns with the concept, while the right heatmap corresponds to the opposing concept. Blue indicates high alignment, red represents low alignment.

Our method successfully identifies distinct conceptual behaviors within the model: when the VLM produces dishonest or unfair responses, token-wise projection scores show clear drops (red regions), in contrast to the consistently high scores observed for honest or fair cases. Such findings provide strong evidence that abstract ethical dimensions like honesty and fairness are internally structured and traceable.

B.3.2 LAT Scans for High-level Representations

To capture how high-level concepts evolve and propagate through the model, we employ Linear Attribution Tomography (LAT) (Zou et al., 2023), which enables layer-wise visualization of conceptual information flow. Fig. 8 shows the resulting LAT scans.

The scans reveal concept-specific propagation patterns. Honesty and fairness exhibit stable trajectories under aligned prompts but greater dispersion and deviation under misaligned ones. Power appears concentrated in later layers, while fearlessness shows early-layer changes. These results are well explained by the SPP framework: concepts propagate through a small set of dominant spectral directions, and the stability observed in aligned cases reflects the spectral gap preservation predicted by Theorem 4.5. The consistency of these representations across modalities further demonstrates the robustness of RepE, and their traceability back to the input is explained by ISLH.