Compression Hacking: A Supplementary Perspective on Informatics Metric of Language Models from Geometric Distortion

Anonymous ACL submission

Abstract

Recently, the concept of "compression as intelligence" has provided a novel informatics metric perspective for language models (LMs), emphasizing that highly structured representations signify the intelligence level of LMs. However, from a geometric standpoint, the word representation space of highly compressed LMs tends to degenerate into a highly anisotropic state, which hinders the LM's ability to comprehend instructions and directly impacts its performance. We found this compression-anisotropy synchronicity is essentially the "***Compression Hacking***" in LM representations, where noise-dominated directions tend to create the illusion of high compression rates by sacrificing spatial uniformity. Based on this, we propose three refined compression metrics by incorporating geometric distortion analysis and integrate them into a self-evaluation pipeline. The refined metrics exhibit strong alignment with the LM's comprehensive capabilities, achieving Spearman correlation coefficients above 0.9, significantly outperforming both the original compression and other internal structure-based metrics. This confirms that compression hacking substantially enhances the informatics interpretation of LMs by incorporating geometric distortion of representations.

1 Introduction

Recently, significant efforts have been devoted to exploring the mechanisms by which language models (LMs) process information internally, driving the development of LM self-evaluation (Wei et al., 2024; Wang et al., 2024a,b) independent of specific tasks and model outputs. The concept of "compression as intelligence" (Sutskever, 2023; Deletang et al., 2023; Chen et al., 2025) has provided a novel Informatics interpretation for LMs, emphasizing that LMs eliminate redundant information through training while their representation spaces typically evolve from disordered to structured states. This property leads to a compression-based evaluation metric for LMs that utilizes differential entropy of representations, aiming to reflect model capabilities with their internal structural organization (Pichler et al., 2022; Zhouyin and Liu, 2023; Li et al., 2025). Existing studies have demonstrated strong alignment between this metric and LM scale (Wei et al., 2024; Li et al., 2025), which we have also empirically validated. However, as evidenced by the intuitive case where 175B GPT-3 (Brown et al., 2020) exhibits inferior overall capabilities compared to 32B Qwen2.5-Instruct (Hui et al., 2024), compression from a purely informatics standpoint, cannot fully align with LM capabilities, especially when comparing models from different families. Therefore, our research motivation is: *Beyond information compression, what other properties should a metric quantify to effectively interpret the LMs' intelligence level, and how should we model the relationships between these properties?*

Relevant studies have shown that differences in model architecture and training paradigms inevitably lead to variations in the geometric structure of representations (Mimno and Thompson, 2017; Gao et al., 2019a; Skean et al., 2025). From a geometric standpoint, we were surprised to observe that LMs with high information compression tend to exhibit representation spaces that degenerate into highly anisotropic, distorted states. Highly anisotropic representations indicate varying sensitivity to semantic changes across different dimensions, which can hinder language models' ability to comprehend instructions and consequently degrade their performance (Demeter et al., 2020; Yu et al., 2022; Rudman and Eickhoff, 2024).

In this study, we quantitatively analyze this compression-anisotropy synchronicity and validate its statistical significance. Through mechanistic analysis, we find that this phenomenon reflects the "***Compression Hacking***" in LM representations, where *noise-dominated directions tend to create*

the illusion of high compression rates by sacrificing spatial uniformity. According to this characteristic, we propose the integration of geometric perspective to refine the information compression metric. Specifically, we introduce the following strategies: (1) a *spectral entropy quantification* compression metric to model the properties of eigenvalue distributions; (2) a *semantic coefficient of variation* to measure anisotropy relative to compression; and (3) a *manifold correction protocol* that uses Principal Component Smoothing (PCS) as an "anisotropy razor" to decouple the influence of anisotropy on compression. These refined metrics are integrated into a self-evaluation pipeline that relies entirely on the LM's internal structure.

Using this framework, we evaluate 18 open-source LMs and conduct meta-evaluations on factuality, reasoning, math, and knowledge tasks to obtain ground-truth capability scores. Extensive experiments demonstrate that the refined metrics exhibit strong alignment with the LM's comprehensive capabilities, achieving Spearman correlation coefficients above 0.9, which significantly outperforms both the original compression and other internal structure-based metrics. This validating the compression hacking substantially enhances the informatics interpretation of LMs by incorporating geometric distortion analysis of representations. The main contributions are summarized as follows:

- We introduce a significant characteristic in LM representations termed "compression hacking", which complements the concept of "compression as intelligence" from the perspective of geometric distortion.
- According to compression hacking, we propose three refinements of compression metrics incorporating geometric insights: spectral entropy quantification, semantic coefficient of variation, and manifold correction protocol.
- The refined metrics exhibit significantly stronger alignment with LM's comprehensive capabilities compared to original compression metric, thereby establishing a task-agnostic self-evaluation perspective for LMs. Anonymous codes available here.

2 Compression Hacking

In this section, we analyze the compression-anisotropy synchronicity in LM representations, where highly compressed LMs tend to exhibit word representations with strong anisotropy. Our investigation proceeds in two stages: First, we quantify both compression and anisotropy metrics by examining the internal structure of LM representations (covariance matrices). We then fit regression curves to model the relationship between anisotropy and compression, verifying it's statistical significance. Second, through mechanistic analysis, we identify the underlying cause of this phenomenon, what we term "compression hacking".

The covariance matrix of LM representations reflects their internal structure. For the hidden states $\mathbf{Z} = \{\mathbf{z}(\boldsymbol{w}) | \boldsymbol{w} \in \mathcal{V}\}$, where $\boldsymbol{w}$ represents a word and $\mathcal{V}$ represents the sample vocabulary space, the construction of the covariance matrix is as formulated in Eq. 1. Here, $\mathbf{z}(\boldsymbol{w}) \in \mathbb{R}^D$ represents the token embeddings, which has been normalized. $\mathbf{Z}$ is a zero-mean matrix.

$$\Sigma_{\mathbf{Z}} = \frac{1}{|\mathcal{V}|}\mathbf{Z}^\top\mathbf{Z} + \alpha\mathbf{I}_D \tag{1}$$

Here, $\Sigma_{\mathbf{Z}} \in \mathbb{R}^{D\times D}$ denotes the covariance matrix, and a regularization term $\alpha\mathbf{I}_D$ is added to ensure it is full rank. The matrix $\Sigma_{\mathbf{Z}}$ is positive definite and can be decomposed using eigenvalue decomposition as $\Sigma_{\mathbf{Z}} = \mathbf{Q}\Lambda\mathbf{Q}^\top$. The eigenvalues from Λ are $\{\lambda_d\}_{d=1}^D$, arranged in descending order by default, and $\{\mathbf{q}_d\}_{d=1}^D$ are the corresponding eigenvectors.

2.1 Preliminary: Differential Entropy based Compression Metric

The compression perspective provides an information-theoretic foundation for LM evaluation, revealing the intrinsic connections between model scale, generalization capability, and data volume, thus offering theoretical guidance for optimizing model design (Pichler et al., 2022; Sutskever, 2023; Deletang et al., 2023; Wei et al., 2024; Chen et al., 2025). Related studies have shown that the differential entropy $\mathcal{H}_{\text{DE}}(\mathbf{Z}) = -\mathbb{E}_{\boldsymbol{w}\sim\mathcal{V}}\mathbf{z}(\boldsymbol{w})\log\mathbf{z}(\boldsymbol{w})$ of LM representations $\mathbf{z}(\boldsymbol{w})$ can reflect their compression capacity (Chen et al., 2023a; Zhouyin and Liu, 2023; Li et al., 2025). Lower differential entropy suggests that the representations formed by nonlinear transformation, which removes redundant information, are closer to optimal coding. These representations exhibit more concentrated distributions and lower uncertainty, reflecting more efficient information compression (Delétang et al.,

2023). Semantic Volume leverages this property to model representation uncertainty (Li et al., 2025).

We thus define compression metric as the negative differential entropy of representations (i.e., $\mathcal{C}_{\text{DE}}(\mathbf{Z}) \stackrel{\text{def}}{=} -\mathcal{H}_{\text{DE}}(\mathbf{Z})$). Since the differential entropy is equivalent to the logdet estimator (Chen et al., 2023a) of their covariance matrix, the compression metric follows the definition in Eq. 2.

$$\mathcal{C}_{\text{DE}}(\mathbf{Z}) \stackrel{\text{def}}{=} -\frac{1}{2}\text{logdet}\left(\Sigma_{\mathbf{Z}}\right) = -\frac{1}{2}\sum_{d=1}^{D}\log\lambda_d \tag{2}$$

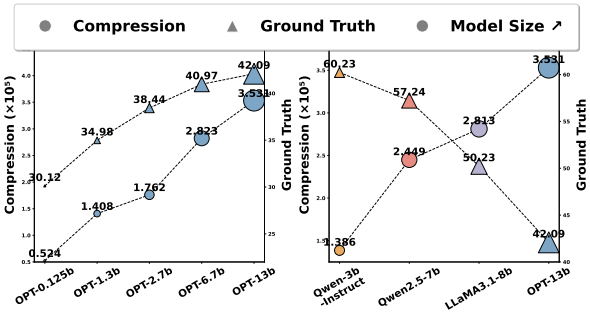

Figure 1: Comparison of compression metrics across different models and their corresponding ground-truth comprehensive capabilities, categorized into intra-family and cross-family comparsions.

We first conducted preliminary exploration to assess whether differential entropy-based compression metrics effectively reflect LM capabilities. Our evaluation included both intra-family (OPT family) and cross-family tests (Qwen2.5-3b-Instruct, Qwen2.5-7b, LLaMA3.1-8b, and OPT-13b), with ground-truth settings following Section 4.1. As shown in Figure 1, we found that compression metrics showed only positive correlations with model scale, consistent with related studies (Wei et al., 2024; Li et al., 2025). However, Figure 1(left) indicates that differential entropy-based compression is effective only for intra-family evaluation, while Figure 1(right) reveals its limited applicability across diverse architectures and training paradigms. These findings prompted our integration of geometric properties into compression analysis.

2.2 Anisotropy: The Geometric Property Correlated with Compression

The anisotropy of language models is a geometric property of representations that reflects the non-uniform distribution of semantics across different directions in the representation space (Ethayarajh, 2019; Cai et al., 2019; Demeter et al., 2020). Highly anisotropic representations hinder LMs' ability to comprehend instructions (Yu et al., 2022; Rudman and Eickhoff, 2024), directly impairing their overall capabilities. We performed principal component analysis to visualize the word representation spaces of the aforementioned four models. As shown in Figure 2, we made the intriguing observation that models with higher compression levels consistently exhibited greater unevenness in their dimensional distributions, namely, higher anisotropy. This suggests a potential synergistic relationship between compression and anisotropy. If we can quantify this relationship and confirm its statistical significance, it could provide valuable guidance for refining compression metrics.

Current tools for qualitatively and quantitatively analyzing the anisotropy of language models mainly rely on similarity computations of representations (Ethayarajh, 2019; Cai et al., 2019; Rudman et al., 2022). However, what we need is an anisotropy metric that can establish a connection with entropy-based information compression. Relevant studies (Arora et al., 2016; Mu and Viswanath, 2018) have shown that the anisotropy measure $\mathcal{A}$ is mathematically defined as formulated in Eq. 3. We aim to extend this measure to relate to the internal structure of representations (eigenvalues of the covariance matrix).

$$\mathcal{A} = \frac{\max_{\|\mathbf{c}\|=1}\mathcal{Z}(\mathbf{c})}{\min_{\|\mathbf{c}\|=1}\mathcal{Z}(\mathbf{c})} \tag{3}$$

where $\mathcal{Z}(\mathbf{c}) = \sum_{\boldsymbol{w}\in\mathcal{V}}\exp\left(\mathbf{c}^\top\mathbf{z}(\boldsymbol{w})\right)$ is the original partition function should approximately be a constant for any unit vector $\mathbf{c}$. $\mathcal{A}$ is a number greater than 1, where larger values indicate stronger anisotropy in the represrntation space. Ideally, this value should be as close to 1 as possible. Considering that $\arg\max_{\|\mathbf{c}\|=1}\mathcal{Z}(\mathbf{c})$ and $\arg\min_{\|\mathbf{c}\|=1}\mathcal{Z}(\mathbf{c})$ do not have closed-form solutions, we attempt to approximate $\mathcal{Z}(\mathbf{c})$ via Taylor expansion as formulated in Eq. 4.

$$\begin{aligned}\mathcal{Z}(\mathbf{c}) = |\mathcal{V}| + \mathbf{1}_{|\mathcal{V}|}^\top\mathbf{Z}\mathbf{c} + \frac{1}{2}\mathbf{c}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{c} \\ + \sum_{m=3}^{\infty}\frac{1}{m!}\sum_{\boldsymbol{w}\in\mathcal{V}}\left(\mathbf{c}^\top\mathbf{z}(\boldsymbol{w})\right)^m\end{aligned} \tag{4}$$

Considering that $\mathbf{Z}$ is zero-mean data, the mean of $\mathbf{z}(\boldsymbol{w})$ is 0. Therefore, the linear term can also be simplified to 0, that is, $\mathbf{1}_{|\mathcal{V}|}^\top\mathbf{Z}\mathbf{c} =$

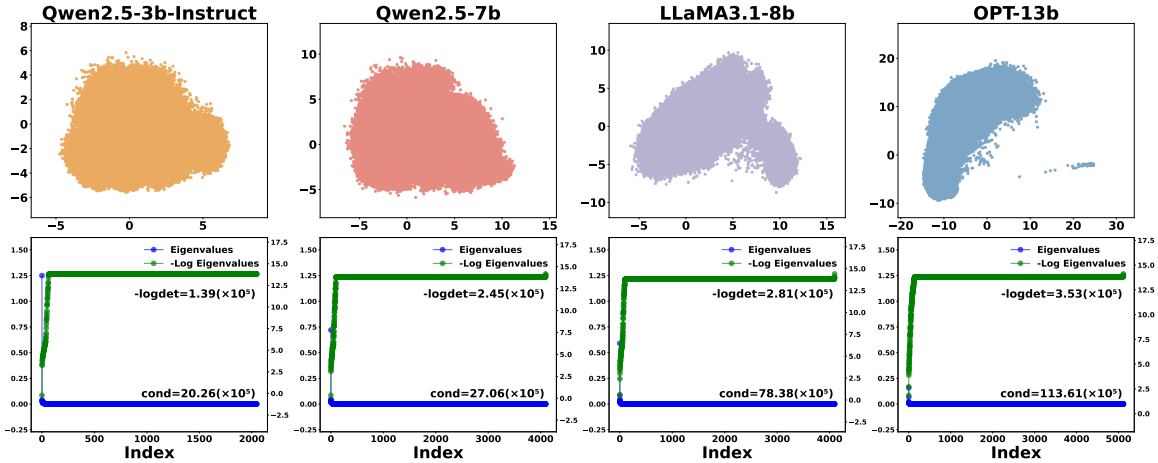

Figure 2: Visualization of distribution of word representations and the eigenvalues across different models.

$\left(\sum_{\boldsymbol{w}\in\mathcal{V}} \mathbf{z}(\boldsymbol{w})\right)^\top \mathbf{c} = \mathbf{0}^\top \mathbf{c} = 0$ which will not affect the relative changes of $\mathcal{Z}(\mathbf{c})$ in different directions. The quadratic term involves the spectral properties of the matrix, whose eigenvalues describe the directional variability of $\mathbf{Z}^\top\mathbf{Z}$, playing a dominant role in the changes of $\mathcal{Z}(\mathbf{c})$ in different directions. Expanding $\mathbf{c}$ in the eigenvector basis, we have $\mathbf{c} = \mathbf{Q}\mathbf{u}$, where $\|\mathbf{u}\| = \|\mathbf{c}\| = 1$ and $\{u_d\}_{d=1}^D$ are the components of $\mathbf{u}$. Based on the eigenvalue decomposition, the calculation of Eq. 5 is made.

$$\mathbf{c}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{c} = (\mathbf{Q}\mathbf{u})^\top \mathbf{Z}^\top\mathbf{Z}\,(\mathbf{Q}\mathbf{u}) = \mathbf{u}^\top\Lambda\mathbf{u} \quad (5)$$

Accordingly, we can further obtain the second-order estimate of $\mathcal{A}$ as formulated in Eq. 6.

$$\begin{aligned}\mathcal{A} &\approx \frac{|\mathcal{V}| + \max_{\|\mathbf{c}\|=1} \frac{1}{2}\mathbf{c}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{c}}{|\mathcal{V}| + \min_{\|\mathbf{c}\|=1} \frac{1}{2}\mathbf{c}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{c}} \\ &= \frac{|\mathcal{V}| + \max_{\|\mathbf{u}\|=1} \frac{1}{2}\sum_d \lambda_d u_d^2}{|\mathcal{V}| + \min_{\|\mathbf{u}\|=1} \frac{1}{2}\sum_d \lambda_d u_d^2}\end{aligned} \quad (6)$$

When the components of the vector $\mathbf{u}$ are entirely concentrated in the direction corresponding to the maximum (minimum) eigenvalue, $\mathbf{u}^\top\Lambda\mathbf{u} = \max_d \lambda_d (\min_d \lambda_d)$. We observed that the anisotropy of the representation can be measured by the condition number of the matrix, as formulated in Eq. 7. The condition number reflects the sensitivity of the covariance matrix and reveals the characteristics of ill-conditioning from an intrinsic structural perspective, making it the first anisotropy metric entirely based on internal structure.

$$\mathcal{A}(\mathbf{Z}) \stackrel{\text{def}}{=} \operatorname{cond}(\Sigma_{\mathbf{Z}}) = \frac{\max_{d=1}^D \lambda_d}{\min_{d=1}^D \lambda_d} \quad (7)$$

2.3 Systematic Analysis

Mechanistic Analysis As shown in Figure 2, by performing eigenvalue decomposition on the covariance matrix of the representations, we discovered a distinctive partitioning phenomenon in the eigenvalues of the LM covariance matrix. The leading principal components exhibit an exponential decay in eigenvalues, effectively condensing the model's core semantic information, while the numerous subsequent minor components demonstrate clustered, nearly constant low eigenvalues, forming spatially anisotropic perturbation sources. Interestingly, when measuring information compression using a negative logarithmic scale, the minor components show dramatically inflated compression metrics due to their infinitesimal original eigenvalues, creating an inverted relationship with the principal component region. This seemingly paradoxical phenomenon actually reveals the compression hacking in model representations, where ***noise-dominated directions tend to create the illusion of high compression rates by sacrificing spatial uniformity***, while in reality this "compression" represents either information loss or noise amplification, with truly effective information compression being exclusively accomplished by the principal components.

Significance Analysis Next, we analyze the significance of compression hacking, which manifests as compression-anisotropy synchronicity. Based on the aforementioned metrics, we calculated the estimates of both compression and anisotropy for instruction representations across four LMs in our preliminary experiments, both of which can be exclusively represented by the eigenvalues of the representation covariance matrix. Given their charac-

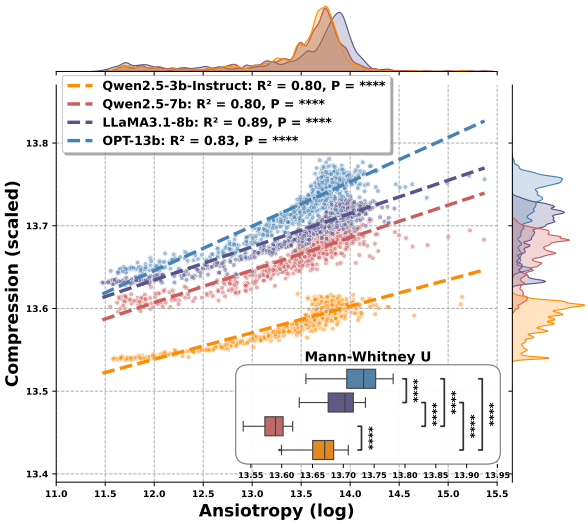

Figure 3: Regression fitting curves of compression versus anisotropy for different models, along with Mann-Whitney U tests between them. Here, **** denotes statistical significance at the 0.01% levels respectively.

teristic patterns, we modeled a linear regression of compression against the logarithmic values of anisotropy, as shown in Figure 3. The regression analysis reveals two key findings through R^2 and p-values: (1) *compression as the dependent variable can be well and significantly explained by anisotropy*, and (2) Mann-Whitney U tests (McKnight and Najab, 2010) confirm *statistically significant differences in regression curves across different models*.

3 Methodology

3.1 Refined Metrics

We have demonstrated that the compression-anisotropy synchronicity caused by compression hacking in LMs is a statistically significant characteristic. This implies that we can develop more comprehensive metrics by jointly considering the compression and anisotropy of representations, as well as modeling their correlation. In this section, we formalize our approach through three strategies:

Spectral Entropy Quantification Figure 2 illustrates that, from the perspective of eigenvalue distribution, the mechanism of compression hacking is that the secondary components causing anisotropy (λ_d) are homologous to the principal components of the compression part ($-\log \lambda_d$). Interesting, spectral entropy (Roy and Vetterli, 2007) precisely models this characteristic, and it is formally equivalent to a compression metric weighted by eigenvalues (Compression (SE)), as formulated in Eq. 8.

$$\mathcal{C}_{\text{SE}}(\mathbf{Z}) \stackrel{\text{def}}{=} -\operatorname{tr}(\Sigma_{\mathbf{Z}} \log \Sigma_{\mathbf{Z}}) = -\sum_{d=1}^{D} \lambda_d \log \lambda_d \tag{8}$$

Semantic Coefficient of Variation Just as compression-anisotropy synchronicity serves as a distinct manifestation of compression hacking, where compression is characterized by the mean of eigenvalue logarithms (reflecting the overall volume of the embedding space (Li et al., 2025)), while anisotropy corresponds to the ratio of extreme eigenvalues (quantifying the variation of semantic embeddings across different dimensions). Thus, we formulate their ratio as the Semantic Coefficient of Variation (Semantic CV) in Eq. 9. This metric accurately characterizes the magnitude of anisotropy relative to information compression in the representation space $\mathbf{Z}$.

$$\mathcal{CV}_{\text{Sem.}}(\mathbf{Z}) \stackrel{\text{def}}{=} \frac{\mathcal{A}(\mathbf{Z})}{\mathcal{C}_{\text{DE}}(\mathbf{Z})} \tag{9}$$

Manifold Correction Protocol Numerous studies have proposed train-free "anisotropy razors" to reduce the anisotropy of representation space in a train-free manner, thereby enhancing representational capacity (Mu and Viswanath, 2018; Su et al., 2021). This inspires us to decouple anisotropy from compression by selecting an appropriate anisotropy razor. Considering the exponential sharp decline in the eigenvalues of principal components corresponding to preceding dimensions due to compression hacking, we propose Principal Component Smoothing (PCS) as an anisotropy razor, inspired by the LW-shrinkage (Ledoit and Wolf, 2004). By setting a smoothing coefficient $\beta \in [0, 1]$ (default value is set to 0.9), we shift the representation space toward principal directions, resulting in a flatter transformed feature spectrum. This transformation is based on the covariance matrix of the representation and is achieved by defining the mapping $\mathcal{T}_{\text{PCS}}$ as formulated in Eq. 10, thereby refining the compression metric Compression (PCS). In Theorem B.2, we prove that under sparse spectrum conditions, the PCS estimator exhibits higher statistical stability than the LW shrinkage.

$$\mathcal{T}_{\text{PCS}}(\Sigma_{\mathbf{Z}}) \stackrel{\text{def}}{=} (1-\beta)\Sigma_{\mathbf{Z}} + \beta \max_{d=1}^{D} \lambda_d \mathbf{I}_D \tag{10}$$

3.2 Evaluation Pipeline

In this section, we integrate the three refined metrics into a unified evaluation framework, which is a

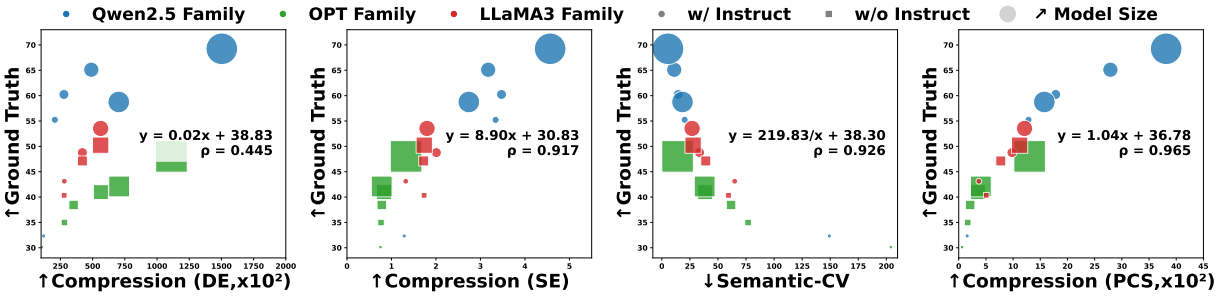

Figure 4: Scatter plots of ground truth values across different models for the four metrics, along with fitted regression equations and Spearman correlation coefficients.

Metric	Global		Qwen2.5-Instruct		OPT		LLaMA3	
	Size	Ground truth	Size	Ground truth	Size	Ground truth	Size	Ground truth
Compression (DE)	0.935	0.445	1.000	0.829	1.000	1.000	0.956	0.886
Compression (SE)	0.430	0.917	0.486	0.714	0.829	0.829	0.598	0.657
Semantic CV	0.805	0.926	0.829	1.000	1.000	1.000	0.956	0.943
Compression (PCS)	0.708	0.965	0.829	1.000	1.000	1.000	0.956	0.943

Table 1: The Spearman correlation coefficients within model groups (Qwen2.5-Instruct, OPT, and LLaMA3 families) and across all models (Global), including the correlations between the four metrics, and both model size (size) and comprehensive capabilities (Ground truth). The gray-highlighted components represent our refined metrics.

task-agnostic pipeline operating purely from a representational perspective. Our evaluation paradigm associates the sampled data batch $\mathcal{B}$ with a decision score $s = \mathcal{F}(\mathcal{B}, f_{\text{LM}})$. The decision function $\mathcal{F}(\cdot)$ operates through two sequential processes: (1) the projection step extracts hidden representations $\mathbf{Z}^{(\boldsymbol{p})} = \mathcal{F}_{\text{Projection}}(\boldsymbol{p}, f_{\text{LM}})$ for each data sample $\boldsymbol{p} \in \mathcal{B}$; (2) the decision step computes the batch-level score $s = \mathbb{E}_{\boldsymbol{p}\sim\mathcal{B}}\text{Metric}(\mathbf{Z}^{\boldsymbol{p}})$ based on the refined metrics. Notably, our dataset requirement specifies that the sample's word representation space should effectively estimate the model's complete word representation space given sufficient sampling, ensuring convergence of our proposed metrics. We discuss the impact of sampling size on metric convergence in Section D.

4 Experiments

In this section, we employ meta-evaluation to investigate whether the refined metrics can achieve strong alignment with the comprehensive capabilities of LMs. This serves to validate whether incorporating the geometric distortion perspective of representations through compression hacking can enhance the informatics interpretation of LMs.

4.1 Setup

Models Since our evaluation focuses on the internal structure of model representations, we evaluated 18 open-source language models from three different model families with varying sizes. These families are the LLaMA3 family (Grattafiori et al., 2024) (LLaMA3.2-1B, LLaMA3.2-1B-Instruct, LLaMA3.2-3B, LLaMA3.2-3B-Instruct, LLaMA3.1-8B, LLaMA3.1-8B-Instruct), Qwen2.5-Instruct family (Hui et al., 2024) (0.5B, 1.5B, 3B, 7B, 14B, 32B), OPT family (Zhang et al., 2022a) (0.125B, 1.3B, 2.7B, 6.7B, 13B, 30B).

Meta Evaluation To evaluate the alignment between our metrics and LM capabilities, we employed meta-evaluation by calculating the Spearman correlation coefficient between human-annotated ground truth benchmarks and our proposed refined informatics metrics. For the meta-evaluation experiments, we selected six benchmark datasets spanning four major domains as ground truth, corresponding to four key dimensions of large language model capabilities: Factuality: TruthfulQA (Lin et al., 2022), FACTOR (Muhlgay et al., 2024), Math: MATH (Hendrycks et al., 2021), Reasoning: CommonsenseQA (Talmor et al., 2019), TheoremQA (Chen et al., 2023b), Knowledge: MMLU (Hendrycks et al., 2020). We use the mean of all benchmark scores as the ground truth for the model's comprehensive evaluation (CE).

Baseline Metrics We selected purely representation-based baseline metrics that operate independently of ground-truth labels and model sampling, encompassing both informatics and geometric perspectives. The informatics

Metric	Property		Factuality		Reasoning		Math	Knowledge	CE
	Info.	Geom.	TruthfulQA	FACTOR	Common.QA	Theo.QA	MATH	MMLU	
Semantic Volume	✓		0.429	0.414	0.441	0.483	0.420	0.409	0.442
Curvature		✓	0.355	0.372	0.342	0.365	0.303	0.309	0.302
Diff-eRank	✓	✓	0.476	0.461	0.494	0.521	0.424	0.452	0.492
Compression(DE)	✓		0.458	0.488	0.481	0.471	0.490	0.471	0.482
Anisotropy		✓	0.715	0.702	0.702	0.792	0.673	0.709	0.701
Compression (SE)	✓	✓	0.895	0.861	0.892	0.921	0.824	0.852	0.912
Semantic CV	✓	✓	0.946	0.905	0.916	0.926	0.857	0.917	0.926
Compression (DE)									
w/ Remove Directions	✓	✓	0.053	0.102	0.042	0.142	0.211	0.093	0.110
w/ Whitening	✓	✓	0.487	0.498	0.502	0.482	0.423	0.456	0.472
w/ LW Shrinkage	✓	✓	0.458	0.488	0.481	0.471	0.490	0.471	0.482
w/ PCS	✓	✓	0.962	0.955	0.923	0.967	0.846	0.923	0.965

Table 2: The Spearman correlation coefficient between the metrics based on the representation properties and the ground truth benchmark, where gray-highlighted components represent refined metrics we proposed.

metrics include Compression (DE) and Semantic Volume (Li et al., 2025), while the geometric metrics consist of Curvature (Hosseini and Fedorenko, 2023) quantifying manifold curvature characteristics, and anisotropy. Diff-eRank (Wei et al., 2024) is the metric that simultaneously models both information compression and geometric structure in language model representations, yet neglecting their direct synergistic relationship.

Baseline Anisotropy Razors In addition to PCS as the anisotropy razor for decoupling anisotropy from compression, we selected three anisotropy razors as baselines. Remove Directions (Mu and Viswanath, 2018) is a post-processing method for eliminating noisy directions. Whitening (Su et al., 2021) eliminates correlations between features through global scaling, normalizing the eigenvalues to have the same mean and variance. LW Shrinkage (Ledoit and Wolf, 2004), on the other hand, adjusts extreme eigenvalues linearly towards the mean via Bayesian shrinkage.

4.2 Main Results

Figure 4 and Table 1 present the regression equations and Spearman correlations among the original compression metric (compression (DE)), our three proposed refined metrics, and comprehensive capabilities as ground truth. The original compression (DE) exhibits strong correlations of 0.935 with model size across all models, reaching 1.000, 1.000, and 0.956 within model families, confirming the high consistency between original compression capability and model scale in language models. However, this metric achieves only 0.445 correlation with comprehensive capabilities in cross-architecture global analysis, maintaining higher correlations (0.829, 1.000, 0.886) only within model families, suggesting model size' applicability for LM capability assessment is confined to homogeneous architectural systems.

Among our refined metrics, compression (SE) shows reduced size correlation (0.430 globally) but achieves 0.917 cross-architecture capability correlation, demonstrating its effectiveness in capturing capability differences across diverse architectures. Both semantic CV and compression (PCS) maintain dual high correlations with size and capabilities within model families while sustaining stable cross-architecture capability correlations (0.926 and 0.965, respectively), with size correlations moderately decreasing to 0.805 and 0.708. This demonstrates that our refined metrics achieve significantly stronger alignment with LMs' comprehensive capabilities compared to the original compression metrics. Through compression hacking, we substantially enhance the informatics interpretation of LMs from the geometric distortion perspective of representations, thereby extending the "compression as intelligence" concept.

4.3 Comparison with Baseline Metrics

Table 2 systematically presents the Spearman correlation coefficients between the ground truth benchmarks, and both the baseline metrics based on internal representations and our refined metrics. The property column identifies whether the metric describes informatics (Info.) or geometric (Geom.) property. Notably, metrics that model only a single property (either informational or geometric property), such as semantic volume, curvature, com-

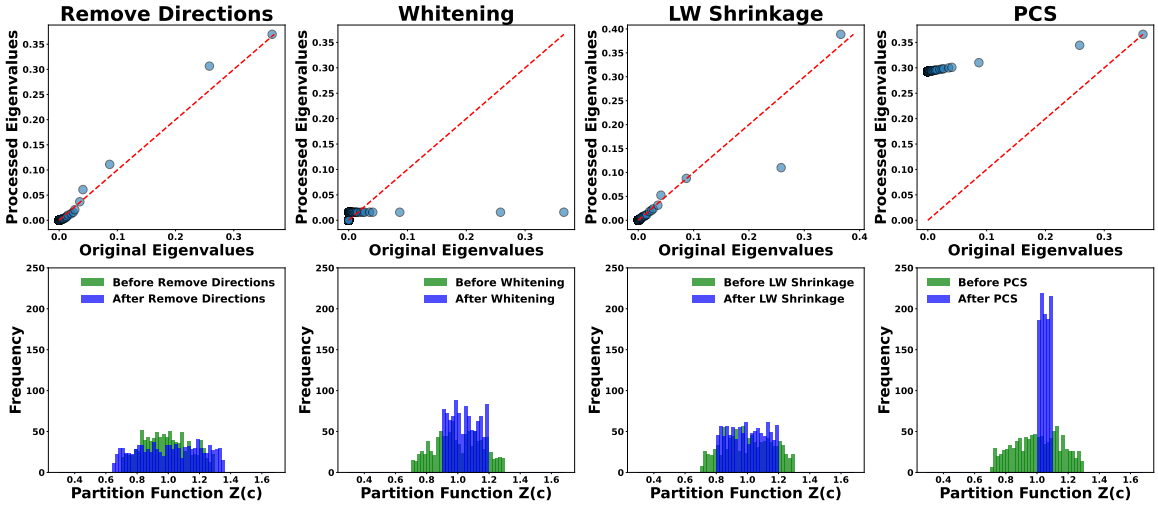

Figure 5: The qqplot of the eigenvalue distribution before and after using different anisotropy razors, and the distribution of the partition function $\mathcal{Z}(\mathbf{c})$.

pression (DE), anisotropy, and their modified versions (w/ remove directions), all exhibit correlation coefficients with the comprehensive score below 0.5. Although Diff-eRank incorporates spectral entropy characteristics, its results still fail to reflect comprehensive capabilities, possibly because this metric focuses on the noise reduction process of knowledge acquisition while neglecting the synergy between information and geometric properties. Experiments show that the compression methods modified by whitening and LW shrinkage, although aiming to decouple anisotropic features, still do not significantly improve capability alignment. It is noteworthy that our refined metrics in Figure 4 demonstrate significant advantages over the baseline metrics.

4.4 Effect of Anisotropy Razors

Table 2 reveals that as "anisotropy razor" methods, remove directions, whitening, and LW shrinkage all fail to effectively improve the reflection of comprehensive capabilities, whereas PCS exhibits a significant improvement. In this section, we investigate the structural changes in representations before and after processing with these anisotropy razors, conducting an in-depth mechanistic analysis of PCS's advantages over other methods.

The qqplot in Figure 5 illustrates the eigenvalue distributions before and after applying these four razors. The first three methods decouple anisotropy while maintaining the linear geometric structure of the data, resulting in eigenvalues that still exhibit distinct partitioning. In contrast, PCS upscales the low-eigenvalue region, ensuring that the corrected compression relies entirely on the contributions of the principal components. The formal method for anisotropy detection involves examining the "self-normalization" property (i.e., $\mathcal{Z}(\mathbf{c})$ tending toward a constant, independent of $\mathbf{c}$) (Mu and Viswanath, 2018). Figure 5 illustrates the distribution of $\mathcal{Z}(\mathbf{c})$ before and after applying different anisotropy razors. We observe that remove directions leads to a more dispersed $\mathcal{Z}(\mathbf{c})$ distribution, increasing anisotropy. This occurs because truncating certain directions causes the remaining ones to spread more extremely. In contrast, whitening, LW shrinkage, and PCS concentrate the $\mathcal{Z}(\mathbf{c})$ distribution. Notably, PCS achieves more pronounced anisotropy elimination than the other methods by rigidly correcting the eigenvalue distribution.

5 Conclusion

We introduce a notable characteristic in language models termed "compression hacking", where the noisy directions in LM representations feign high compression rates by sacrificing spatial uniformity, thereby distorting information compression metrics. Through spectral entropy quantification, semantic coefficient of variation, and a manifold correction protocol based on principal component smoothing, we refine the compression measurement framework. Extensive experiments on 18 mainstream language models demonstrate that the refined metrics achieve strong alignment with models' actual capabilities. These results prove that incorporating the geometric distortion perspective through compression hacking significantly enhances the informatics interpretation of LMs.

6 Limitations

In fact, the metrics we propose still have broader application scenarios worth exploring. For instance, practical techniques such as pruning, quantization, and distillation could potentially benefit from these indicators that reveal internal redundancies. Our proposed metrics help better identify compressible components in models without causing significant information loss. We anticipate that these refined metrics may open new avenues for future research, exploring how such internal representation indicators can be applied to various potential scenarios.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

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

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2019. Isotropy in the contextual embedding space: Clusters and manifolds. In *International conference on learning representations*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2023a. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Jun Chen, Yong Fang, Ashish Khisti, Ayfer Özgür, and Nir Shlezinger. 2025. Information compression in the ai era: Recent advances and future challenges. *IEEE Journal on Selected Areas in Communications*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023b. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Gregoire Deletang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. 2023. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. 2023. Language modeling is compression. *arXiv preprint arXiv:2309.10668*.

David Demeter, Gregory Kimmel, and Doug Downey. 2020. Stolen probability: A structural weakness of neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2191–2197.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Wikimedia Foundation. 2025. Wikimedia downloads. Accessed: 2025-03-02.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019a. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019b. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 6894. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Eghbal Hosseini and Evelina Fedorenko. 2023. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, 36:43918–43930.

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

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837.

Olivier Ledoit and Michael Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Xiaomin Li, Zhou Yu, Ziji Zhang, Yingying Zhuang, Swair Shah, and Anurag Beniwal. 2025. Semantic volume: Quantifying and detecting both external and internal uncertainty in llms. *arXiv preprint arXiv:2502.21239*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66.

Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. 2022. A differential entropy estimator for training neural networks. In *International Conference on Machine Learning*, pages 17691–17715. PMLR.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.

William Rudman and Carsten Eickhoff. 2024. Stable anisotropic regularization. In *ICLR*.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. Isoscore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339.

Yutaka Sasaki et al. 2007. The truth of the f-measure. *Teach tutor mater*, 1(5):1–5.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Ilya Sutskever. 2023. Stronger compressors find more shared structure. The Ilya's Talk. Talk.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Zhiquan Tan, Lai Wei, Jindong Wang, Xing Xie, and Weiran Huang. 2024. Can i understand what i create? self-knowledge evaluation of large language models. *arXiv preprint arXiv:2406.06140*.

Yiming Wang, Pei Zhang, Baosong Yang, Derek Wong, Zhuosheng Zhang, and Rui Wang. 2024a. Embedding trajectory for out-of-distribution detection in mathematical reasoning. *Advances in Neural Information Processing Systems*, 37:42965–42999.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F Wong, and Rui Wang. 2024b. Latent space chain-of-embedding enables output-free llm self-evaluation. *arXiv preprint arXiv:2410.13640*.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. 2024. Diff-erank: A novel rank-based metric for evaluating large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 2022. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29–45.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022b. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11730–11738.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Zhanghao Zhouyin and Ding Liu. 2023. Understanding neural networks withlogarithm determinant entropy estimator.

A Related Work and Further Analysis

A.1 Evaluation of Language Models

The evaluation of language models is currently in a state of rapid iterative development, encompassing a variety of tasks, datasets, and benchmarks (Celikyilmaz et al., 2020; Zheng et al., 2023; Tan et al., 2024). Traditional evaluation metrics such as accuracy, F1-score (Sasaki et al., 2007), BLEU (Sellam et al., 2020), and ROUGE (Lin, 2004) focus on comparing model predictions with annotated labels in downstream tasks. Other metrics like perplexity and cross-entropy loss do not rely on annotated labels and are computed solely based on input text. However, these methods primarily emphasize external evaluation based on model predictions.

Recently, significant efforts have been devoted to exploring the mechanisms by which language models (LMs) process information internally, driving the development of LM self-evaluation (Wei et al., 2024; Wang et al., 2024a,b) independent of specific tasks and model outputs. The concept of "compression as intelligence" has provided an information-theoretic internal evaluation perspective for language models, highlighting that the acquisition of world knowledge by language models is a denoising process (Sutskever, 2023; Deletang et al., 2023; Wei et al., 2024; Chen et al., 2025). Differential entropy of representations, as a classical information-theoretic measure, effectively quantifies the internal uncertainty of language models (Chen et al., 2023a; Zhouyin and Liu, 2023). Semantic volume (Li et al., 2025) leverages representation-level differential entropy-aware compression metrics to offer a novel perspective for language model evaluation. However, related work has found that such compression can only model the scale of language models and fails to align with their capabilities (Wei et al., 2024; Li et al., 2025).

We introduces the concept of compression hacking in language model representations, where the noisy directions of LM representations sacrifice spatial uniformity to feign high compression rates. This implies that we can refine the information compression perspective by considering the geometric distortions in the language model's representation space.

A.2 Anisotropy of Language Models

Anisotropy The anisotropy of language models reflects the geometric properties of the contextual embedding space. Related studies have observed that during sampling, the spatial embeddings of negative samples exhibit anisotropy, which describes how vectors are distributed within the contextual space (Mimno and Thompson, 2017; Ethayarajh, 2019). The researchers found that most vectors occupy a relatively narrow cone within the space, and that vectors within this cone tend to have high cosine similarity (Gao et al., 2019b). Demeter pointed out that using softmax introduces structural weaknesses in the representation space, leading to bias, a common issue in language models (Demeter et al., 2020). To better quantify the anisotropy of LMs, related work has identified isolated clusters and low-dimensional manifolds in the contextual embedding space, introducing tools for their qualitative and quantitative analysis (Ethayarajh, 2019; Cai et al., 2019; Rudman et al., 2022). However, these tools are mainly based on similarity calculations of embedded representations. What is needed instead is an anisotropy metric that can establish a connection with entropy based compression metric.

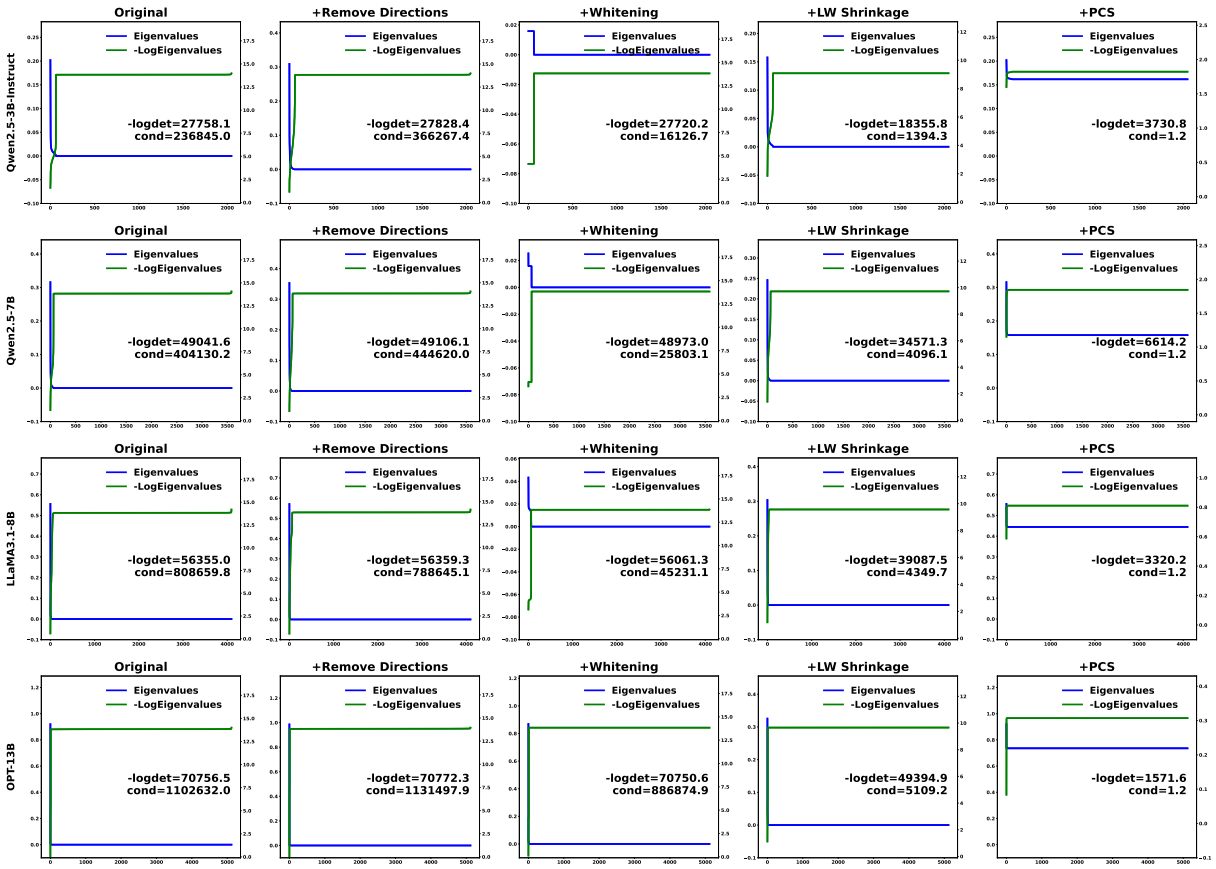

Figure 6: The eigenvalues and their negative logarithmic distributions of different models' representations before and after processing with different anisotropy razors.

Anisotropy Razors To mitigate anisotropy in language models, existing research has proposed various solutions. Contrastive learning has emerged as a powerful tool for obtaining effective sentence representations, effectively reducing anisotropy by increasing the spatial distance between positive and negative samples (Gao et al., 2021; Zhang et al., 2022b; Jiang et al., 2022). In this work, we employ post-processing methods applied directly to the representation space as baseline approaches for the anisotropy razor:

- **Remove Directions** (Mu and Viswanath, 2018): First, subtract the common mean vector of all word vectors to eliminate global bias; then remove the top high-variance principal component directions via Principal Component Analysis (PCA). This process enhances semantic feature discriminability by eliminating non-semantic common information from word vectors, making the word space distribution more isotropic.

- **Whitening** (Su et al., 2021): Zero-center the representations and transform the covariance matrix into an identity matrix, forcing the embedding distribution toward isotropy.

- **LW Shrinkage** (Ledoit and Wolf, 2004): Linearly shrink the sample covariance matrix toward the diagonal matrices to reduce noise interference in high-dimensional data, yielding more stable covariance matrix estimates. This operation mitigates excessive sensitivity in specific directions, promoting isotropic feature distributions.

These training-free paradigms provide references for decoupling anisotropy from compression. However, these methods maintain the linear geometric structure of the data, with eigenvalues still exhibiting consistent partitioning behavior. Figure 6 demonstrates the distribution changes in eigenvalues and their negative logarithms after applying these baseline anisotropy razor post-processing methods. The results show that the distributions after Remove Directions, Whitening, and LW-Shrinkage treatments retain their original forms, leaving cross-model relationships of the modified compression metrics relatively un-

changed. Consequently, we propose principal component smoothing to force eigenvalues toward dominant features. As shown in Figure 6, this approach induces significant changes in eigenvalue distributions.

B Statistical Properties of Principal Component Smoothing

Lemma B.1 (Asymptotic Optimality of Ledoit–Wolf Shrinkage (Ledoit and Wolf, 2004))**.** *Let $\Sigma \in \mathbb{R}^{D\times D}$ be the population covariance matrix and $\Sigma_{\mathbf{Z}} = \frac{1}{|\mathcal{V}|}\mathbf{Z}^\top\mathbf{Z}$ the sample covariance. The Ledoit-Wolf estimator*

$$\hat{\Sigma}_{LW} = (1-\beta_{LW})\Sigma_{\mathbf{Z}} + \beta_{LW}\mu\mathbf{I}, \quad \mu = \frac{1}{D}\operatorname{tr}(\Sigma_{\mathbf{Z}}) \tag{11}$$

attains minimal MSE when the shrinkage intensity satisfies $\beta_{LW} \asymp \frac{1}{|\mathcal{V}|}$. Under general covariance structures (without spectral sparsity), this yields asymptotic MSE:

$$MSE(\hat{\Sigma}_{LW}) \asymp \mathcal{O}\left(\frac{D}{|\mathcal{V}|}\right) \tag{12}$$

Theorem B.2 (Statistical Stability of the Principal Component Smoothing Estimator)**.** *Assume the true covariance matrix Σ has a dominant eigenvalue $\lambda_1^* = \max_d \lambda_d \gg \lambda_d^* \quad (d \geq 2)$, i.e., spectral sparsity holds. Define the improved shrinkage estimator as:*

$$\hat{\Sigma}_{PCS} = (1-\beta_{PCS})\Sigma_{\mathbf{Z}} + \beta_{PCS}\lambda_1\mathbf{I}, \beta_{PCS} \asymp \mathcal{O}\left(\frac{1}{\sqrt{|\mathcal{V}|}}\right) \tag{13}$$

where λ_1 is the largest eigenvalue of the sample covariance matrix $\Sigma_{\mathbf{Z}}$ and satisfies $\lambda_1 \xrightarrow{|\mathcal{V}|} \lambda_1^*$ in probability. When the sample size $|\mathcal{V}|$ is sufficiently large,

$$\mathrm{MSE}(\hat{\Sigma}_{PCS}) < \mathrm{MSE}(\hat{\Sigma}_{LW}) \tag{14}$$

Proof. We commence by analyzing the mean squared error (MSE) structure of covariance matrix estimators. Let $\|\cdot\|_F$ denote the Frobenius norm, the MSE decomposes into bias and variance components:

$$\mathrm{MSE}(\hat{\Sigma}) = \underbrace{\left\|\mathbb{E}[\hat{\Sigma}] - \Sigma\right\|_F^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left\|\hat{\Sigma} - \mathbb{E}[\hat{\Sigma}]\right\|_F^2\right]}_{\text{Variance}}. \tag{15}$$

For the Ledoit-Wolf estimator $\hat{\Sigma}_{\mathrm{LW}} = (1-\beta_{\mathrm{LW}})\Sigma_{\mathbf{Z}} + \beta_{\mathrm{LW}}\mu\mathbf{I}$, under spectral sparsity $\lambda_1^* \gg \sum_{d=2}^{D}\lambda_d^*/D$, the shrinkage target $\mu \approx \lambda_1^*/D$ creates dominant bias from the leading eigenvalue:

$$\begin{aligned}\mathrm{Bias}_{\mathrm{LW}}^2 &\approx \beta_{\mathrm{LW}}^2\|\Sigma - \mu\mathbf{I}\|_F^2 \\ &= \beta_{\mathrm{LW}}^2\left[(\lambda_1^*-\mu)^2 + \sum_{d=2}^{D}(\lambda_d^*-\mu)^2\right] \\ &\asymp \beta_{\mathrm{LW}}^2(\lambda_1^*)^2\left(1-\tfrac{1}{D}\right)^2\end{aligned} \tag{16}$$

According to lemma B.1, the variance term inherits from sample covariance matrix with dimension scaling:

$$\mathrm{Variance}_{\mathrm{LW}} \approx (1-\beta_{\mathrm{LW}})^2 \cdot \mathcal{O}\left(\tfrac{D^2}{|\mathcal{V}|}\right) \asymp \mathcal{O}\left(\tfrac{D^2}{|\mathcal{V}|}\right) \tag{17}$$

where the $\mathcal{O}(D^2/|\mathcal{V}|)$ scaling comes from concentration of sample covariance in high dimensions.

For our eigenvalue-shrinkage estimator $\hat{\Sigma}_{\mathrm{PCS}} = (1-\beta_{\mathrm{PCS}})\Sigma_{\mathbf{Z}} + \beta_{\mathrm{PCS}}\lambda_1\mathbf{I}$, the preserved leading eigenvalue estimation $\lambda_1 \xrightarrow{p} \lambda_1^*$ fundamentally alters the bias-variance tradeoff. The bias now originates from minor eigenvalues:

$$\mathrm{Bias}_{\mathrm{PCS}}^2 = \beta_{\mathrm{PCS}}^2\sum_{d=2}^{D}(\lambda_d^*-\lambda_1^*)^2 \asymp \beta_{\mathrm{PCS}}^2(D-1)(\lambda_1^*)^2 \tag{18}$$

where the last approximation uses $\lambda_d^* \ll \lambda_1^*$ from spectral sparsity. The variance term splits into two parts:

$$\begin{aligned}\mathrm{Variance}_{\mathrm{PCS}} = (1-\beta_{\mathrm{PCS}})^2 &\underbrace{\mathrm{Var}\left(\sum_{d=2}^{D}\lambda_d\right)}_{\asymp\mathcal{O}\left(\frac{(D-1)\lambda_1^{*2}}{|\mathcal{V}|}\right)} \\ + \beta_{\mathrm{PCS}}^2 &\underbrace{\mathrm{Var}(\lambda_1)}_{\asymp\mathcal{O}\left(\frac{\lambda_1^{*2}}{|\mathcal{V}|}\right)}\end{aligned} \tag{19}$$

With optimal shrinkage intensity $\beta_{\mathrm{PCS}} = \mathcal{O}(1/\sqrt{|\mathcal{V}|})$, the dominant variance term becomes:

$$\mathrm{Variance}_{\mathrm{PCS}} \asymp \mathcal{O}\left(\tfrac{(D-1)\lambda_1^{*2}}{|\mathcal{V}|}\right). \tag{20}$$

The MSE comparison reveals fundamental differences in scaling laws. For $\hat{\Sigma}_{\mathrm{LW}}$ with $\beta_{\mathrm{LW}} = \mathcal{O}(1/|\mathcal{V}|)$:

$$\mathrm{MSE}(\hat{\Sigma}_{\mathrm{LW}}) \asymp \underbrace{\mathcal{O}\left(\tfrac{\lambda_1^{*2}}{|\mathcal{V}|^2}\right)}_{\text{Bias}^2} + \underbrace{\mathcal{O}\left(\tfrac{D^2}{|\mathcal{V}|}\right)}_{\text{Variance}}. \tag{21}$$

Model	R²	p-value
LLaMA3.1-8B	0.89	****
LLaMA3.1-8B-Instruct	0.79	***
LLaMA3.2-1B	0.80	****
LLaMA3.2-1B-Instruct	0.78	***
LLaMA3.2-3B	0.89	****
LLaMA3.2-3B-Instruct	0.77	****
OPT-0.125B	0.88	****
OPT-1.3B	0.76	****
OPT-2.7B	0.66	**
OPT-6.7B	0.91	****
OPT-13B	0.80	***
OPT-30B	0.83	****
Qwen2.5-0.5B-Instruct	0.81	****
Qwen2.5-1.5B-Instruct	0.86	***
Qwen2.5-3B-Instruct	0.80	****
Qwen2.5-7B-Instruct	0.79	****
Qwen2.5-14B-Instruct	0.85	****
Qwen2.5-32B-Instruct	0.83	****

Table 3: The R² and p-values of the compression-anisotropy regression fitting curves across different models, where, **, ***, and **** denote statistical significance at the 1%, 0.1%, and 0.01% levels respectively.

For $\hat{\Sigma}_{\text{PCS}}$ with dimension-adaptive shrinkage:

$$\text{MSE}(\hat{\Sigma}_{\text{PCS}}) \asymp \underbrace{\mathcal{O}\left(\frac{(D-1)\lambda_1^{*2}}{|\mathcal{V}|}\right)}_{\text{Bias}^2} + \underbrace{\mathcal{O}\left(\frac{(D-1)\lambda_1^{*2}}{|\mathcal{V}|}\right)}_{\text{Variance}}. \tag{22}$$

When $|\mathcal{V}| \to \infty$, the $\mathcal{O}(1/|\mathcal{V}|)$ terms dominate $\mathcal{O}(1/|\mathcal{V}|^2)$. Under spectral sparsity $\lambda_1^* \gg \lambda_d^*$ ($d \geq 2$), the improvement ratio becomes:

$$\frac{\text{MSE}(\hat{\Sigma}_{\text{PCS}})}{\text{MSE}(\hat{\Sigma}_{\text{LW}})} \asymp \frac{D\lambda_1^{*2}/|\mathcal{V}|}{D^2/|\mathcal{V}|} = \frac{\lambda_1^{*2}}{D} \ll 1, \tag{23}$$

where the inequality follows from $\lambda_1^{*2}/D \leq (\sum_{d=1}^{D} \lambda_d^*)^2/D^2$ by Cauchy-Schwarz. ■

C Significance Analysis

Our evaluation results presented in Table 3 demonstrate a strong and statistically significant relationship between compression and anisotropy across the 18 open-source language models examined. The high R² values (ranging from 0.7 to 0.9 for most models) indicate that linguistic anisotropy accounts for a substantial proportion of the observed compression phenomena. Furthermore, the compression-anisotropy synchronization proves statistically significant at stringent confidence levels (p<0.001 or p<0.01) for the majority of models. These robust and consistent findings across diverse architectures provide compelling empirical evidence that compression hacking is not merely an artifact but rather an intrinsic and fundamental characteristic of language model representations, revealing important insights about their underlying geometric properties.

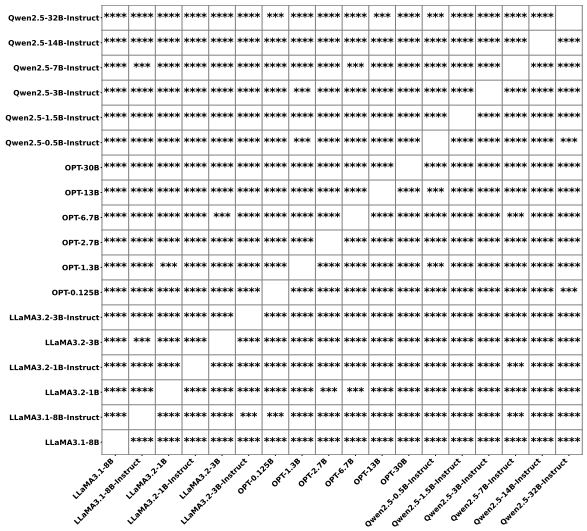

Figure 7: The Mann-Whitney U tests of compression-anisotropy regression fitting between different models, where, ***, and **** denote statistical significance at the 0.1%, and 0.01% levels respectively.

Figure 7 presents the Mann-Whitney U test results for compression-anisotropy regression fitting across different models. Our analysis reveals that the differences between most model pairs achieve statistical significance at rigorous levels. These statistically significant variations in compression-anisotropy fitting curves demonstrate that the information compression metric, when adjusted for compression hacking effects, can effectively capture meaningful distinctions in model capabilities. This finding provides empirical validation that our refined compression-based evaluation framework offers discriminative power for comparing performance differences across language model architectures.

D Implementation Details of the Evaluation Pipeline

For the projection dataset, we primarily collected 1,000 data samples from the pretraining corpus (Wiki (Foundation, 2025)) and the instruction-tuning dataset (Dolly-15k (Conover et al., 2023)) to derive projection data. By sampling the word repre-

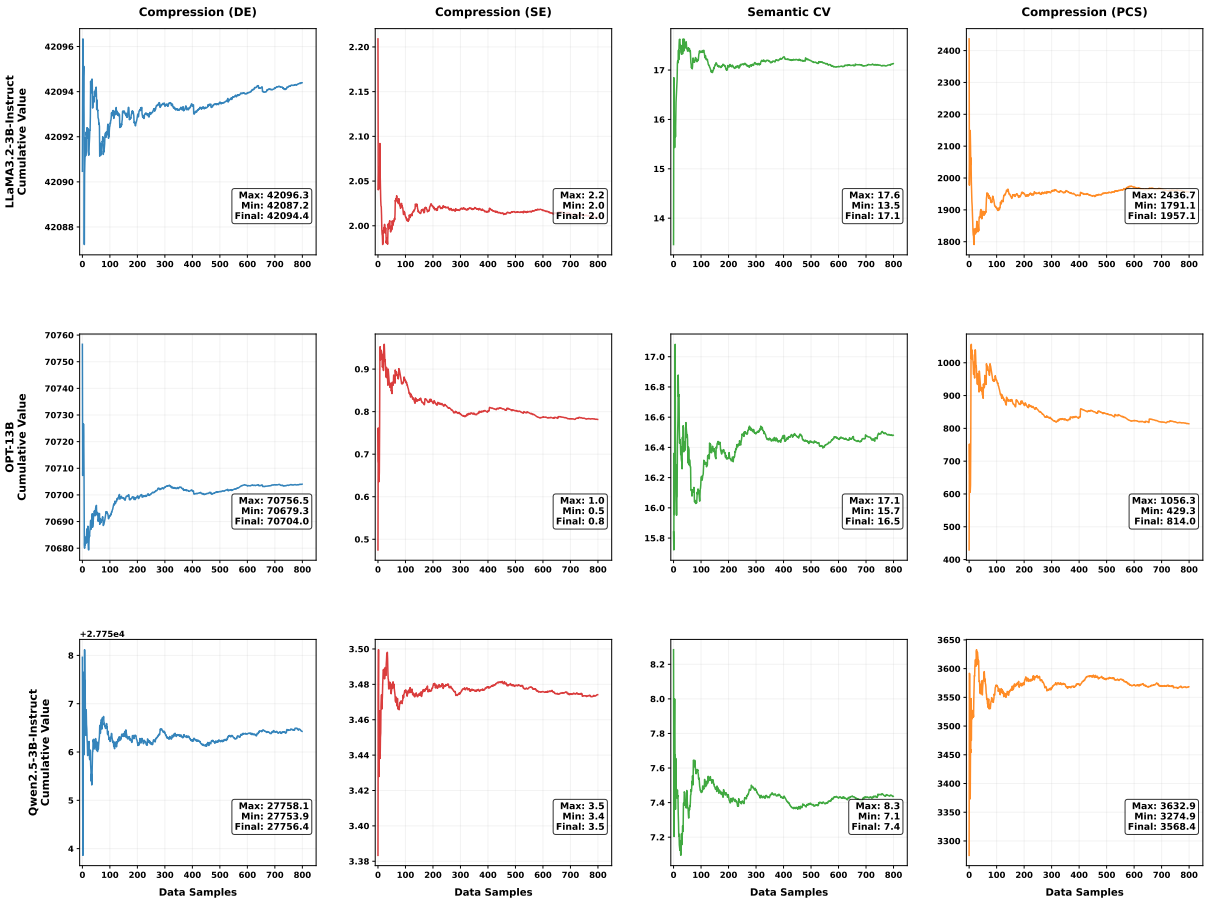

Figure 8: The cumulative expected values of different metrics as the number of samples increases.

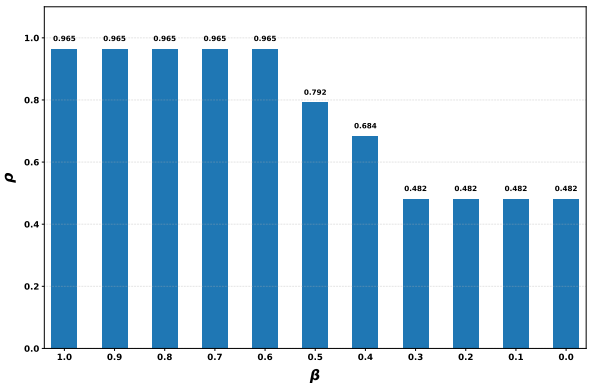

Figure 9: The correlation coefficients between compression (PCS) and ground truth under different smoothing coefficient.

sentations of these data points, we aim to estimate the full model's representation space, ensuring the convergence of our metrics. Our pipeline defaults to sampling 800 data samples. Figure 8 illustrates the cumulative expected values of different metrics as the number of samples increases. We observe that all metrics converge relatively early to stable values, demonstrating that our refined metrics enable robust evaluation based on the provided projection dataset.

For the hyperparameter α that ensures full-rank covariance matrices, we selected 10^{-8}. Regarding the smoothing coefficient (β) for principal component smoothing, we determined the interval $[0.6, 1]$ to be appropriate. Figure 8 illustrates how different choices of principal component smoothing coefficients affect the compression (PCS). It can be observed that when β falls within $[0.6, 1]$, the results maintain strong correlation with the ground truth. This occurs because the principal directions already dominate the compression computation. As the smoothing coefficient decreases, noise directions gradually regain prominence in the compression calculation.