COMPRESSION HACKING: TOWARDS A UNIFIED INFORMATICS AND GEOMETRIC PERSPECTIVE ON INTRINSIC EVALUATION OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, the concept of "compression as intelligence" has provided a novel informatics metric perspective for large language models (LLMs), emphasizing that highly structured representations signify the intelligence level of LLMs. However, from a geometric standpoint, the representation space of highly compressed LLMs tends to degenerate into a highly anisotropic state, which hinders the LLM's ability to comprehend instructions and directly impacts its performance. We found this compression-anisotropy synchronicity is essentially the "***Compression Hacking***" in LLM 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 an intrinsic-evaluation framework. The refined metrics exhibit strong alignment with the LLM'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 information-theoretic perspective on LLMs by incorporating geometric distortion of representations.

1 INTRODUCTION

Recently, significant efforts have been devoted to exploring the mechanisms by which large language models (LLMs) process information internally, driving the development of LLMs intrinsic-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 information-theoretic perspective for LLMs, emphasizing that LLMs 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 LLMs that utilizes a Gaussian differential-entropy surrogate of representations, aiming to reflect model capabilities with their internal structural organization (Zhouyin & Liu, 2023; Chen et al., 2023; Li et al., 2025). Existing studies have demonstrated strong alignment between this metric and LLM 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 LLMs 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 LLMs' 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 & Thompson, 2017; Gao et al., 2019a; Skean et al., 2025). From a geometric standpoint, we were surprised to observe that LLMs 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 LLMs' ability to comprehend instructions and consequently degrade their performance (Demeter et al., 2020; Yu et al., 2022; Rudman & Eickhoff, 2024).

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

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 LLM representations, where *noise-dominated directions tend to create the illusion of high compression rates by sacrificing spatial uniformity* (Mu et al., 2017; Zhang et al., 2024; Jha & Reagen, 2025). 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) to decouple the influence of anisotropy on compression. These refined metrics are integrated into an intrinsic-evaluation framework that relies entirely on the LLMs' internal structure.

Using this framework, we evaluate 18 open-source LLMs and conduct meta-evaluations on factuality, reasoning, math, knowledge, and agentic capability tasks to obtain ground-truth capability scores. Extensive experiments demonstrate that the refined metrics exhibit strong alignment with the LLMs comprehensive capabilities, achieving Spearman correlation coefficients above 0.9, which significantly outperform both the original compression and other internal structure-based metrics. This validates that compression hacking substantially enhances the information-theoretic perspective on LLMs by incorporating geometric distortion analysis of representations.

In conclusion, the main contributions are summarized as follows:

- We introduce a significant characteristic in LLM 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 LLMs' intelligence compared to the original compression metric, thereby establishing a task-agnostic intrinsic-evaluation perspective for LLMs.

2 Compression Hacking in Large Language Models

In this section, we reveal the "compression hacking" phenomenon in LLMs, where highly compressed LLMs tend to exhibit strong anisotropy in their in-context representations. First, we quantify both compression and anisotropy metrics by examining the internal structure of LLM representations. Subsequently, we fit regression curves to model the relationship between anisotropy and compression, and verify its statistical significance. Second, through mechanistic analysis, we uncover the underlying cause of this phenomenon, which we term "compression hacking".

The covariance matrix of LLM 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 $\Sigma_{\mathbf{Z}} = \frac{1}{|\mathcal{V}|}\mathbf{Z}^\top\mathbf{Z}$. Here, $\mathbf{z}(\boldsymbol{w}) \in \mathbb{R}^D$ represents the token hidden state, which has been normalized. $\mathbf{Z}$ is a zero-mean matrix, $\Sigma_{\mathbf{Z}} \in \mathbb{R}^{D\times D}$ denotes the covariance matrix, which 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: LogDet-based Compression Metric

The compression perspective provides an information-theoretic foundation for LLMs 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 use LogDet-based entropy surrogates of LLM representations to characterize their compression capacity (Chen et al., 2023; Zhouyin & Liu, 2023; Li et al., 2025). Lower surrogate 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).

Following prior LogDet-based studies (Zhouyin & Liu, 2023; Chen et al., 2023; Li et al., 2025), we use the entropy of the covariance-matched Gaussian as a tractable second-order surrogate for the representation distribution. For a Gaussian vector, differential entropy is determined by the covariance log-determinant up to an additive constant (Lemma G.1). Its covariance-dependent component gives the compression score in Equation 1.

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

We first conducted preliminary exploration to assess whether LogDet-based compression metrics effectively reflect LLMs 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 observe that compression metrics exhibit only a positive correlation with model scale. This is consistent with the findings of previous research on LLMs (Wei et al., 2024; Li et al., 2025) and LLMs-based agents (He et al., 2025). However, while Figure 1 (left) suggests that LogDet-based compression is effective for intra-family evaluation, Figure 1 (right) reveals its limitations across diverse architectures and training paradigms. A case in point is OPT-13B, which, despite its larger scale, demonstrates inferior overall capabilities compared to Qwen2.5-3B-Instruct. These findings prompted our integration of geometric properties into compression analysis.

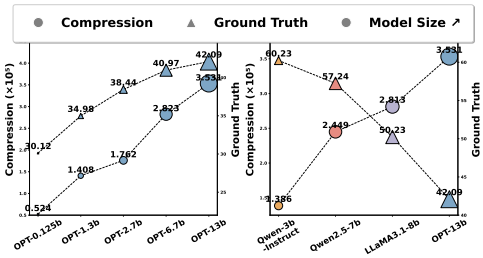


Figure 1: Comparison of compression metrics across different models and their corresponding ground-truth comprehensive capabilities, categorized into intra-family and cross-family comparisons.

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 LLMs' ability to comprehend instructions (Yu et al., 2022; Rudman & Eickhoff, 2024), directly impairing their overall capabilities. We performed principal component analysis to visualize the in-context 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 & Viswanath, 2018) have shown that the anisotropy measure $\mathcal{A}(\mathbf{c})$ is mathematically defined as formulated in Equation 2. We aim to extend this measure to relate to the internal structure of representations (eigenvalues of the covariance matrix).

$$\mathcal{A}(\mathbf{c}) = \frac{\max_{\|\mathbf{c}\|=1}\sum_{\boldsymbol{w}\in\mathcal{V}}\exp\left(\mathbf{c}^{\top}\mathbf{z}(\boldsymbol{w})\right)}{\min_{\|\mathbf{c}\|=1}\sum_{\boldsymbol{w}\in\mathcal{V}}\exp\left(\mathbf{c}^{\top}\mathbf{z}(\boldsymbol{w})\right)} \tag{2}$$

where $\mathcal{Z}(\mathbf{c}) \stackrel{\mathrm{def}}{=} \sum_{\boldsymbol{w}\in\mathcal{V}}\exp\left(\mathbf{c}^{\top}\mathbf{z}(\boldsymbol{w})\right)$ is the original partition function, should approximately be a constant for any unit vector $\mathbf{c}$. $\mathcal{A}(\mathbf{c})$ is a number greater than 1, where larger values indicate

stronger anisotropy in the representation 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 Equation 3.

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

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, $\mathbf{1}_{|\mathcal{V}|}^{\top}\mathbf{Z}\mathbf{c} = \left(\sum_{\boldsymbol{w}\in\mathcal{V}}\mathbf{z}(\boldsymbol{w})\right)^{\top}\mathbf{c} = \mathbf{0}^{\top}\mathbf{c} = 0$ 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 $\mathbf{Z}^{\top}\mathbf{Z}$, thereby relating the directional changes of $\mathcal{Z}(\mathbf{c})$ to the spectrum. Expanding $\mathbf{c}$ in the eigenvector basis, we have $\mathbf{c} = \mathbf{Q}\mathbf{u}$, where $\|\mathbf{u}\| = \|\mathbf{c}\| = 1$. Based on the eigenvalue decomposition, the calculation of Equation 4 is made.

$$\mathbf{c}^{\top}\mathbf{Z}^{\top}\mathbf{Z}\mathbf{c} = (\mathbf{Q}\mathbf{u})^{\top}\mathbf{Z}^{\top}\mathbf{Z}(\mathbf{Q}\mathbf{u}) = \mathbf{u}^{\top}\Lambda\mathbf{u} \tag{4}$$

Accordingly, we can further obtain the second-order estimate of $\mathcal{A}$ as formulated in Equation 5.

$$\mathcal{A}(\mathbf{c}) \approx \frac{|\mathcal{V}| + \max_{\|\mathbf{c}\|=1}\{\mathbf{1}_{|\mathcal{V}|}^{\top}\mathbf{Z}\mathbf{c} + \frac{1}{2}\mathbf{c}^{\top}\mathbf{Z}^{\top}\mathbf{Z}\mathbf{c}\}}{|\mathcal{V}| + \min_{\|\mathbf{c}\|=1}\{\mathbf{1}_{|\mathcal{V}|}^{\top}\mathbf{Z}\mathbf{c} + \frac{1}{2}\mathbf{c}^{\top}\mathbf{Z}^{\top}\mathbf{Z}\mathbf{c}\}} = \frac{|\mathcal{V}| + \max_{\|\mathbf{u}\|=1}\frac{1}{2}\mathbf{u}^{\top}\Lambda\mathbf{u}}{|\mathcal{V}| + \min_{\|\mathbf{u}\|=1}\frac{1}{2}\mathbf{u}^{\top}\Lambda\mathbf{u}} \tag{5}$$

According to the properties of the Rayleigh quotient, the quadratic extrema in Equation 5 are determined by the maximum and minimum eigenvalues. Thus, the second-order estimate establishes a relationship between partition-function anisotropy and the spectral extremes. Motivated by this relationship, we use the condition number in Equation 6 as a spectrum-based anisotropy measure. It quantifies the relative separation between the spectral extremes and the ill-conditioning of the covariance matrix from an intrinsic structural perspective.

$$\mathcal{A}(\mathbf{Z}) \overset{\text{def}}{=} \operatorname{cond}(\Sigma_{\mathbf{Z}}) = \frac{\max_{d=1}^{D}\lambda_d}{\min_{d=1}^{D}\lambda_d} \tag{6}$$

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 LLMs covariance matrix. Further combining the results in Table 1, it can be observed that the eigenvalues of the top principal components exhibit an exponential decay, effectively concentrating the model's core semantic information (Mu et al., 2017; Zhang et al., 2024). Specifically, a very small number of principal components (less than 1% of the total dimensions) are sufficient to explain 95% of the variance. In contrast, the subsequent vast majority of minor components display clustered and nearly constant low eigenvalues, serving as the source of perturbation for anisotropy (Jha & Reagen, 2025).

Table 1: Eigenvalue statistics and principal components of representation spaces across different LLMs. We report the maximum and minimum eigenvalues, the condition number, and the k for cumulative explained variance ratio of 95% (i.e. number of principal components (k) required to explain 95% of the variance).

Model	$\max_d \lambda_d$	$\min_d \lambda_d$	$\frac{\max_d \lambda_d}{\min_d \lambda_d}$	k for 95% CEVR
Qwen2.5-3b-Instruct	1.25	6.06×10^{-6}	2.06×10^{5}	20 (dim:2048)
Qwen2.5-7b	0.72	2.61×10^{-6}	2.75×10^{5}	33 (dim:4096)
LLaMA3.1-8b	0.59	7.42×10^{-7}	7.95×10^{5}	32 (dim:4096)
OPT-13b	0.16	1.36×10^{-7}	1.17×10^{6}	22 (dim:5120)

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 (Mu et al., 2017; Zhang et al., 2024; Jha & Reagen, 2025).

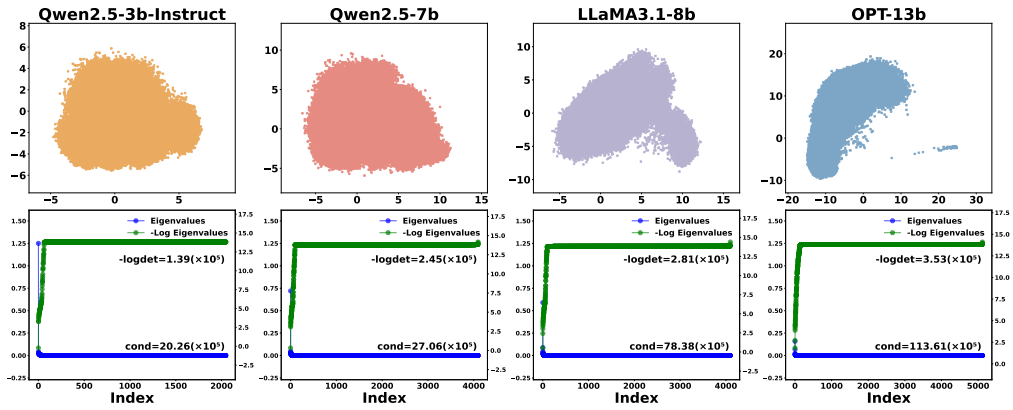


Figure 2: Visualization of distribution of in-context representations and the eigenvalues across different models. The horizontal axis of the trend plot represents the representation feature index.

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 LLMs in our preliminary experiments, both of which can be exclusively represented by the eigenvalues of the representation covariance matrix. Given their characteristic 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) the linear regression yields an $R^2 > 0.8$ and p-values significant at the 0.01% level, indicating that *compression as the dependent variable is well and significantly explained by anisotropy*; (2) the p-values in the Mann-Whitney U tests (McKnight & Najab, 2010) reach significance at the 0.01% level, confirming *statistically significant differences in regression curves across different models*. The conclusions presented above illustrate that at the statistical granularity, the property of compression hacking exhibits significant effects on individual models, while notable discrepancies arise across different models.

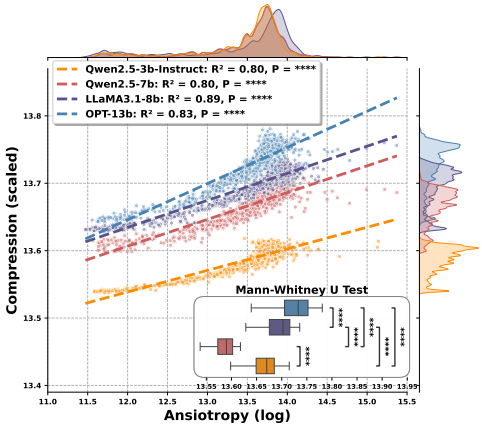


Figure 3: Regression fitting curves of compression versus anisotropy for different models, along with Mann-Whitney U tests between them. Here, **** denotes statistical significance at the 0.01% level.

3 Methodology

3.1 Refined Metrics

We have demonstrated that the compression–anisotropy synchronicity caused by compression hacking in LLMs 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. These three strategies provide parallel characterizations of compression–anisotropy synchronicity from different perspectives, and possess consistent reference value as intrinsic indicators of model capability.

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 ($-\log \lambda_d$). Interestingly, spectral entropy (Roy & Vetterli, 2007) precisely models this characteristic, and it

is formally equivalent to a compression metric weighted by eigenvalues (Compression (SE)), as formulated in Equation 7. Notably, when the covariance matrix is positive semi-definite, this refined metric possesses a closed-form connection with the Effective Rank (Zhang et al., 2023). The Effective Rank intuitively represents the effective number of dimensions active within the data representations, thereby quantifying the actual information capacity of the model in the representation space, rather than its mere algebraic rank.

$$\mathcal{C}_{\text{SE}}(\mathbf{Z}) \stackrel{\text{def}}{=} -\operatorname{tr}(\Sigma_{\mathbf{Z}} \log \Sigma_{\mathbf{Z}}) = -\sum_{d=1}^{D} \lambda_d \log \lambda_d \tag{7}$$

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 representation space (Li et al., 2025)), while anisotropy corresponds to the ratio of extreme eigenvalues (quantifying the variation of semantic representations across different dimensions). Thus, we formulate their ratio as the Semantic Coefficient of Variation (Semantic CV) in Equation 8. This metric accurately characterizes the magnitude of anisotropy relative to information compression in the representation space $\mathbf{Z}$.

$$\mathcal{CV}_{\text{Sem.}}(\mathbf{Z}) \stackrel{\text{def}}{=} \frac{\mathcal{A}(\mathbf{Z})}{\mathcal{C}_{\text{DE}}(\mathbf{Z})} \tag{8}$$

Manifold Correction Protocol. Numerous studies have proposed train-free anisotropy decoupling methods to reduce the anisotropy of representation space in a train-free manner, thereby enhancing representational capacity (Mu & Viswanath, 2018; Su et al., 2021). This inspires us to decouple anisotropy from compression by selecting an appropriate method. 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) for anisotropy decoupling, inspired by the LW-shrinkage (Ledoit & 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 Equation 9, thereby refining the compression metric Compression (PCS). Rather than comparing PCS and LW as covariance estimators in terms of Frobenius-risk, Proposition G.2 formalizes the role of PCS in our framework: under sparse-spectrum representations, PCS deterministically raises the floor of tail eigenvalues and reduces spectral ill-conditioning, thereby stabilizing the log-det-based compression score.

$$\mathcal{T}_{\text{PCS}}(\Sigma_{\mathbf{Z}}) \stackrel{\text{def}}{=} (1-\beta)\Sigma_{\mathbf{Z}} + \beta \max_{d=1}^{D} \lambda_d \mathbf{I}_D \tag{9}$$

3.2 Evaluation Framework

In this section, we integrate the three refined metrics into a unified evaluation framework, which is a task-agnostic framework operating purely from a representational perspective. Our evaluation paradigm associates the sampled data batch $\mathcal{B}$ with a decision score $s = \mathcal{F}(\mathcal{B}, f_{\text{LLM}})$. To underscore the task-agnostic nature of our metrics, the framework employs representative foundational corpora to activate the representation space: a pre-training corpus (Wiki (Foundation, 2025)) and an instruction-tuning dataset (Dolly-15k (Conover et al., 2023)). The decision function $\mathcal{F}(\cdot)$ operates through two sequential processes: (1) the projection step extracts hidden representations $\mathbf{Z}^{\boldsymbol{p}} = \mathcal{F}_{\text{Projection}}(\boldsymbol{p}, f_{\text{LLM}})$ for each data sample $\boldsymbol{p} \in \mathcal{B}$. By default, we use the hidden states from the last Transformer layer.; (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, the in-context representation space activated by the sampled data should effectively estimate the model's complete representation space, thereby ensuring the convergence of our proposed metrics. A detailed sensitivity analysis of the framework hyperparameters is available in Section E.

4 Experiments

In this section, we employ meta-evaluation to investigate whether the refined metrics can achieve strong alignment with the intelligence of LLMs. This serves to validate whether incorporating the

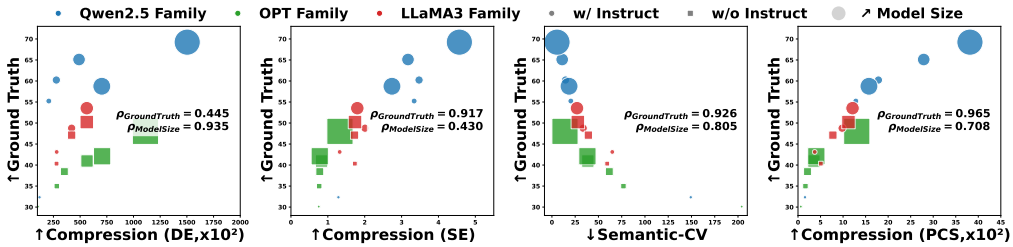


Figure 4: Scatter plots of ground-truth comprehensive evaluation (CE) scores against the four metrics, together with fitted regression equations and Spearman correlations.

geometric distortion perspective of representations through compression hacking can enhance the information-theoretic perspective on LLMs.

4.1 Setup

Models. Considering that our framework relies on eliciting in-context representations, we evaluated 18 open-source LLMs 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 LLMs 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 eight benchmark datasets spanning eight major domains as ground truth: Factuality: TruthfulQA (Lin et al., 2022), Instruction following: IF-Eval (Zhou et al., 2023), Reasoning: CommonsenseQA (Talmor et al., 2019), Knowledge: MMLU (Hendrycks et al., 2020), Math: MATH (Hendrycks et al., 2021), Coding: HumanEval (Chen et al., 2021) Agent: GAIA (Mialon et al., 2023), Context learning: CL-Bench (Dou et al., 2026). We apply the Rank-Sum Ratio (RSR) method to all benchmark scores to obtain the ground truth for the model's comprehensive evaluation (CE); its complete construction is provided in Appendix B. To make larger correlations consistently mean better capability alignment, we direction-align lower-is-better metrics before computing Spearman correlation. In particular, for Semantic CV we report $\rho(1/\text{Semantic CV}, \text{CE})$.

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 metrics include Compression (DE) and Semantic Volume (Li et al., 2025). The geometric metrics consist of Curvature (Hosseini & Fedorenko, 2023), which quantifies manifold curvature characteristics; and anisotropy, which can be measured by condition number (Cond) and explained variance ratio (EVR). Diff-eRank (Wei et al., 2024) is the metric that simultaneously models both information compression and geometric structure in LLM representations, yet neglecting their direct synergistic relationship.

Baseline Anisotropy Decoupling Methods. In addition to PCS for decoupling anisotropy from compression, we selected three anisotropy razors as baselines. Remove Directions (Mu & 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 & Wolf, 2004), on the other hand, adjusts extreme eigenvalues linearly towards the mean via Bayesian shrinkage.

4.2 Main Results

Figure 4 summarizes the relationship between the four compression-related metrics and the ground-truth comprehensive capability scores across all evaluated models. The original LogDet-based metric, Compression (DE), shows limited capability alignment, yielding a Spearman correlation of 0.445. As illustrated in Figure 4, its variation is more consistent with model scale than with capability differences

Table 2: Spearman correlations between representation-based metrics and ground-truth benchmarks, where gray-highlighted rows denote our refined metrics.

Metric	Property		Factuality	Instruction Following	Knowledge	Reasoning	Math	Coding	Agent	Context Learning	CE
	Info.	Geom.									
Semantic Volume	✓		0.429	0.421	0.409	0.483	0.420	0.421	0.401	0.441	0.442
Curvature		✓	0.355	0.365	0.309	0.365	0.303	0.365	0.314	0.342	0.302
Diff-eRank	✓	✓	0.476	0.429	0.452	0.521	0.424	0.481	0.405	0.494	0.492
Compression(DE)	✓		0.458	0.488	0.471	0.471	0.490	0.319	0.521	0.481	0.445
Anisotropy (EVR)		✓	0.707	0.709	0.709	0.744	0.651	0.523	0.673	0.702	0.693
Anisotropy (Cond)		✓	0.715	0.709	0.709	0.792	0.673	0.598	0.673	0.702	0.701
Compression (SE)	✓	✓	0.895	0.861	0.852	0.921	0.824	0.861	0.861	0.892	**0.917**
Semantic CV	✓	✓	0.946	0.962	0.917	0.926	0.857	0.875	0.832	0.916	**0.926**
Compression (DE)											
w/ Remove Directions	✓	✓	0.053	0.134	0.093	0.142	0.211	0.072	0.204	0.042	0.110
w/ Whitening	✓	✓	0.487	0.468	0.456	0.445	0.423	0.431	0.471	0.502	0.472
w/ LW Shrinkage	✓	✓	0.458	0.458	0.471	0.471	0.490	0.251	0.458	0.481	0.445
w/ PCS	✓	✓	0.962	0.955	0.923	0.967	0.846	0.875	0.831	0.923	**0.965**

across model families, suggesting that compression defined purely from the LogDet entropy surrogate does not adequately support cross-family intrinsic evaluation. By contrast, the three refined metrics all exhibit substantially stronger alignment with capability. Compression (SE) improves the correlation to 0.917 by weighting spectral contributions according to eigenvalue magnitude. Semantic CV further reaches a direction-aligned correlation of 0.926 by explicitly relating anisotropy to compression; its raw signed correlation is -0.926, as expected because lower Semantic CV indicates stronger capability. Compression (PCS) achieves the strongest overall alignment, with a Spearman correlation of 0.965, and shows the clearest monotonic trend in Figure 4. Overall, these results indicate that incorporating geometric distortion yields a more capability-relevant characterization of representation compression.

To examine the robustness of these results, Figure 5 reports the distributions of leave-one-model-out Spearman correlations. Specifically, for each metric, we remove one model at a time and recompute the correlation on the remaining 17 models. The full-set correlations are 0.445, 0.917, 0.926, and 0.965 for DE, SE, direction-aligned Semantic CV, and PCS, respectively. Their mean LOO correlations are 0.444, 0.916, 0.925, and 0.964. Thus, every refined metric remains above DE in every held-out case, showing that the central comparison does not depend on any single model. Compression (PCS) is particularly stable, with LOO correlations ranging only from 0.958 to 0.983. Semantic CV and Compression (SE) show the same qualitative pattern, with substantially higher and less variable correlations than DE.

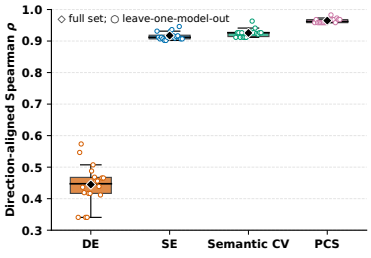


Figure 5: Direction-aligned Spearman correlations under leave-one-model-out (LOO) analysis. Open circles show the 18 LOO values and black diamonds show the full-set correlations. Semantic CV is aligned by rank reversal using its reciprocal.

We use LOO here as a sensitivity diagnostic rather than treating the 18 highly overlapping LOO values as independent samples. Accordingly, Figure 5 reports the full distributions without pairwise significance bars. Figures 4 and 5 therefore provide complementary evidence about the strength of full-set alignment and its sensitivity to removing one evaluated model. Detailed values are provided in Section F.

4.3 Task-level Results & 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. Compared to other metrics, our three proposed metrics maintain significantly higher consistency with the ground-truth capabilities across all evaluation dimensions. In particular, they demonstrate a high degree of alignment in the evaluation of Factuality, Instruction Following, Knowledge, Reasoning, and Context Learning tasks. Notably, metrics that model only a single property (either informational or geometric property), such as semantic volume, curvature, compres-

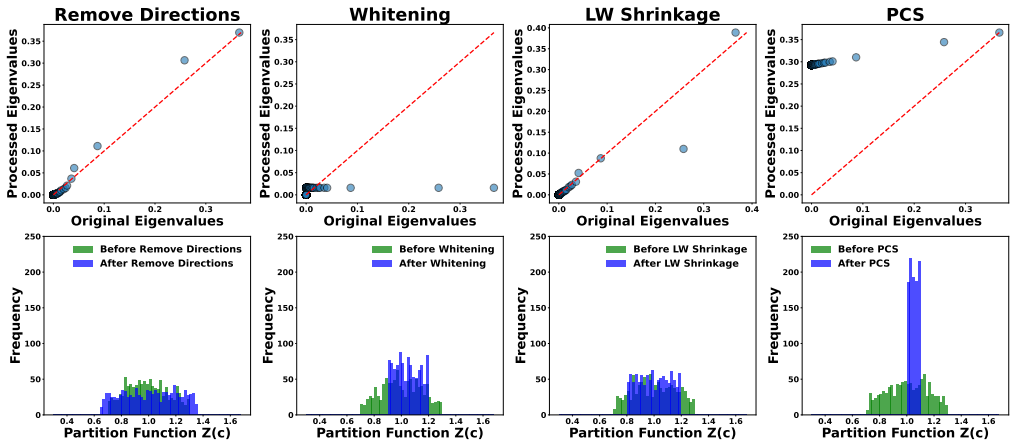


Figure 6: The qqplot of the eigenvalue distribution before and after using different anisotropy decoupling methods, and the distribution of the partition function $\mathcal{Z}(\mathbf{c})$.

sion (DE), anisotropy, and their modified versions (w/ remove directions), all exhibit correlation coefficients with the comprehensive score below 0.8. 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 anisotropy, 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 Decoupling Methods

Table 2 reveals that as anisotropy decoupling 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 decoupling methods, conducting an in-depth mechanistic analysis of PCS's advantages over other methods.

The qqplot in Figure 6 illustrates the eigenvalue distributions before and after applying these four methods. 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 & Viswanath, 2018). Figure 6 illustrates the distribution of $\mathcal{Z}(\mathbf{c})$ before and after applying different methods. 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 LLMs termed "compression hacking", where the noisy directions in LLM 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 LLMs demonstrate that the refined metrics achieve strong alignment with the models' actual capabilities. These results show that incorporating the geometric distortion perspective through compression hacking significantly enhances the information-theoretic perspective on LLMs.

AI USE STATEMENT

In this work, we used generative AI tools for language translation. We have not used generative AI tools for other tasks with required disclosure (e.g., generating synthetic data, formulating mathematical claims, or interpreting results), and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for literature discovery and language editing. We have reviewed all AI-assisted work. We manually verified the relevance of the discovered literature and carefully checked the translated and edited text for accuracy and nuance. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The ethical considerations of our work mainly concern data and model use. We use publicly available datasets and openly released model checkpoints for representation analysis and evaluation. All resources are used solely for academic research in accordance with their licenses and terms of use. We do not collect private data, solicit user data, or conduct experiments involving human participants.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed metrics and evaluation framework in Section 3. The evaluated models, projection datasets, benchmark datasets, baseline metrics, and meta-evaluation protocol are described in Section 4, particularly Section 4.1. Appendix B details the construction of the comprehensive evaluation score, while Appendices E and F report the hyperparameter sensitivity, cross-layer analysis, cross-family results, and leave-one-model-out analysis.

REFERENCES

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

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.

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

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

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

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob

McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

Peter Christen, David J Hand, and Nishadi Kirielle. A review of the f-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3):1–24, 2023.

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

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. Language modeling is compression. *arXiv preprint arXiv:2309.10668*, 2023.

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. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*, 2023.

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

Shihan Dou, Jiazheng Zhang, Jianxiang Zang, Yunbo Tao, Weikang Zhou, Haoxiang Jia, Shichun Liu, Yuming Yang, Zhiheng Xi, Shenxi Wu, et al. Multi-programming language sandbox for llms. *arXiv preprint arXiv:2410.23074*, 2024.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, et al. Cl-bench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*, 2026.

Kawin Ethayarajh. 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, 2019.

Wikimedia Foundation. Wikimedia downloads, 2025. URL `https://dumps.wikimedia.org/`. Accessed: 2025-03-02.

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

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894. Association for Computational Linguistics, 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 e-prints*, pp. arXiv–2407, 2024.

Shizhe He, Avanika Narayan, Ishan S Khare, Scott W Linderman, Christopher Ré, and Dan Biderman. An information theoretic perspective on agentic system design. *arXiv preprint arXiv:2512.21720*, 2025.

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

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

Eghbal Hosseini and Evelina Fedorenko. 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, 2023.

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.

Nandan Kumar Jha and Brandon Reagen. Spectral scaling laws in language models: How effectively do feed-forward networks use their latent space? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35059–35070. Association for Computational Linguistics, 2025.

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

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

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

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

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

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

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

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

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

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. Representing sentences as low-rank subspaces. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 629–634. Association for Computational Linguistics, 2017.

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

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, 2020.

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

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

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

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

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

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

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 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)*, pp. 4149–4158, 2019.

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

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020.

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

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

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

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 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)*, pp. 29–45, 2022.

Gaifan Zhang, Yi Zhou, and Danushka Bollegala. Evaluating unsupervised dimensionality reduction methods for pretrained sentence embeddings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6530–6543. ELRA and ICCL, 2024.

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

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

Yifan Zhang, Zhiquan Tan, Jingqin Yang, Weiran Huang, and Yang Yuan. Matrix information theory for self-supervised learning. *arXiv preprint arXiv:2305.17326*, 2023.

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

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

A Limitations & Broader Impacts

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 may help identify compressible components in models without causing significant information loss. However, two limitations remain. First, although our analysis reveals a strong correlation between compression and geometric distortion, we do not yet provide a fully causal explanation of this relationship. Second, whether these refined metrics can provide actionable guidance for LLM training, rather than post hoc evaluation alone, remains to be explored. We anticipate that these metrics may open new avenues for future research on the use of internal representation indicators in both model evaluation and model optimization. More broadly, as a foundational study of intrinsic evaluation, the societal impact of this work is mainly indirect, primarily through its potential influence on model evaluation and optimization.

B Construction of the Comprehensive Evaluation Score

We construct CE using the Rank-Sum Ratio (RSR), a nonparametric rank-based aggregation method. We first rank the $N = 18$ models separately within each of the $K = 8$ external benchmarks, after orienting every benchmark so that a higher score indicates better performance. The best model receives the highest rank, and tied observations receive their average rank. All eight benchmarks are weighted equally. Let R_{ij} denote the rank of model i on benchmark j. The comprehensive score is

$$\mathrm{CE}_i = 100 \times \frac{\sum_{j=1}^{K} R_{ij}}{KN}, \qquad K = 8,\ N = 18. \tag{10}$$

Because RSR aggregates within-benchmark ranks rather than raw scores, benchmarks reported on different scales, such as accuracy and pass@1, cannot dominate CE solely because of their numerical range. CE uses only the eight external benchmarks listed in Section 4.1; none of the intrinsic metrics proposed in this paper enters its construction. We then compute the Spearman correlation between CE and each intrinsic metric, applying the direction alignment described in Section 4.1 when lower metric values indicate stronger capability.

C Related Work and Further Analysis

C.1 Evaluation of Large Language Models

The evaluation of large language models is evolving rapidly, spanning a wide range 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 (Christen et al., 2023), BLEU (Sellam et al., 2020), and ROUGE (Lin, 2004) primarily assess model outputs by comparing predictions against annotated references in downstream tasks. Other commonly used metrics, such as perplexity and cross-entropy loss, do not require labeled outputs, but they still remain tied to the model's behavior on input text. Overall, these approaches mainly emphasize external evaluation based on observable predictions.

More recently, growing attention has been devoted to understanding how LLMs process information internally, giving rise to intrinsic evaluation methods that do not depend directly on downstream task labels or output judgments (Wei et al., 2024; Wang et al., 2024a;b; Dou et al., 2024). Compared with external evaluation, this line of research aims to characterize model quality through internal representations, structural organization, and latent processing dynamics. Our work belongs to this intrinsic-evaluation setting: rather than assessing the correctness of model outputs on a particular benchmark, we study whether internal representational properties can serve as reliable indicators of model capability.

C.2 Information-Theoretic Perspective on Large Language Models

Information theory provides an important foundation for interpreting deep learning systems. In the context of large language models, this line of work has developed from early studies of representation-level information content to more recent efforts on compression-based intrinsic evaluation. Prior

studies estimated the information encoded in hidden representations using probing-style methods and also investigated the lossless compression limits of LLMs through tools such as arithmetic coding (Pimentel et al., 2020; He et al., 2025; Voita & Titov, 2020).

Within this broader line of research, the idea of "compression as intelligence" has motivated an information-theoretic perspective on model evaluation, suggesting that model scale, generalization, and data efficiency are closely tied to representation compression (Pichler et al., 2022; Sutskever, 2023; Deletang et al., 2023; Wei et al., 2024; Chen et al., 2025). Under this view, differential entropy serves as a natural measure of representational uncertainty, and lower entropy is interpreted as reflecting more concentrated and efficient coding (Chen et al., 2023; Zhouyin & Liu, 2023; Delétang et al., 2023). Semantic Volume further extends this perspective by using differential-entropy-aware compression to characterize uncertainty in LLM representations (Li et al., 2025).

However, existing evidence suggests that compression-based metrics, when used alone, are more strongly aligned with model scale than with model capability (Wei et al., 2024; Li et al., 2025). This limitation has also been observed in LLM-based agents (He et al., 2025). In our setting, we observe the same pattern: pure compression metrics tend to increase with model size, yet fail to reliably distinguish capability differences across architectures and training paradigms. A representative example is that OPT-13B, despite its larger scale, can still underperform Qwen2.5-3B-Instruct in overall capability. This limitation motivates our central hypothesis: to make compression-based intrinsic evaluation more capability-relevant, one must account not only for information concentration, but also for the geometric distortion of the representation space. This is precisely the perspective behind our notion of compression hacking.

C.3 Representation-based Intrinsic Metrics

Recent work on intrinsic evaluation has moved beyond general analyses of internal representations and proposed concrete metrics for assessing LLMs directly from their hidden states, without relying on downstream labels or output judgments. These metrics can be roughly grouped by the representation property they emphasize, including information-oriented metrics, geometry-oriented metrics, and metrics that partially combine the two.

Information-oriented Metrics. A representative line of work measures the uncertainty or compactness of hidden representations from an information-theoretic perspective. In particular, LogDet-based compression metrics quantify representation compression through a Gaussian differential-entropy surrogate of the covariance matrix, based on the view that more concentrated representations indicate stronger compression ability (Zhouyin & Liu, 2023; Chen et al., 2023; Li et al., 2025). Semantic Volume further extends this perspective by characterizing uncertainty in LLM representations with a differential-entropy-aware formulation (Li et al., 2025). These metrics provide natural proxies for representation compactness, but mainly focus on concentration itself.

Geometry-oriented Metrics. Another line of work studies the geometric organization of representations. Curvature measures whether representations evolve along straighter or more structured manifolds (Hosseini & Fedorenko, 2023). A related direction focuses on anisotropy, which describes the non-uniform utilization of directions in the representation space (Ethayarajh, 2019; Rudman & Eickhoff, 2024; Rudman et al., 2022). In practice, anisotropy can be quantified from the covariance spectrum in different ways. The condition number (Cond) measures spectral ill-conditioning through the ratio between extreme eigenvalues (Arora et al., 2016; Mu & Viswanath, 2018), while explained variance ratio (EVR) reflects how strongly variance is concentrated in a small number of dominant directions. These metrics are useful for revealing representational degeneration, but do not directly characterize compression.

Hybrid Metrics. Some recent work has started to connect information and geometry in a single metric. Diff-eRank is a representative example, incorporating spectrum-related characteristics to jointly describe structural organization and information reduction in hidden representations (Wei et al., 2024). Compared with metrics that model only one property, such methods provide a richer view of internal structure. However, they still treat compression and geometry mainly as parallel descriptors, rather than explicitly modeling their interaction.

Overall, existing representation-based intrinsic metrics provide useful but still partial views of LLM internal structure. Information-based metrics emphasize concentration, geometry-based metrics em-

phasize spatial organization, and hybrid metrics begin to bridge the two. Our work differs in explicitly modeling their coupled relationship. Specifically, we argue that strong measured compression may co-occur with severe anisotropy, producing *compression hacking*. This motivates our refined metrics, which incorporate geometric distortion into compression-based intrinsic evaluation in a unified way.

C.4 Representation Anisotropy of Language Models

Anisotropy. The anisotropy of language models reflects the geometric properties of the contextual representation space. Related studies have observed that during sampling, the spatial representations of negative samples exhibit anisotropy, which describes how vectors are distributed within the contextual space (Mimno & 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 LLMs, related work has identified isolated clusters and low-dimensional manifolds in the contextual representation 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.

Anisotropy Decoupling Methods. 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 anisotropy decoupling method:

- **Remove Directions** (Mu & 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 and whiten the representations, forcing the representation distribution toward isotropy.
- **LW Shrinkage** (Ledoit & 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 7 demonstrates the distribution changes in eigenvalues and their negative logarithms after applying these baseline anisotropy decoupling 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 unchanged. Consequently, we propose principal component smoothing to force eigenvalues toward dominant features. As shown in Figure 7, this approach induces significant changes in eigenvalue distributions.

D Significance Analysis of Compression Hacking

Our evaluation results presented in Table 3 demonstrate a strong and statistically significant relationship between compression and anisotropy across the 18 open-source LLMs examined. The high R^2 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 ($p < 0.001$

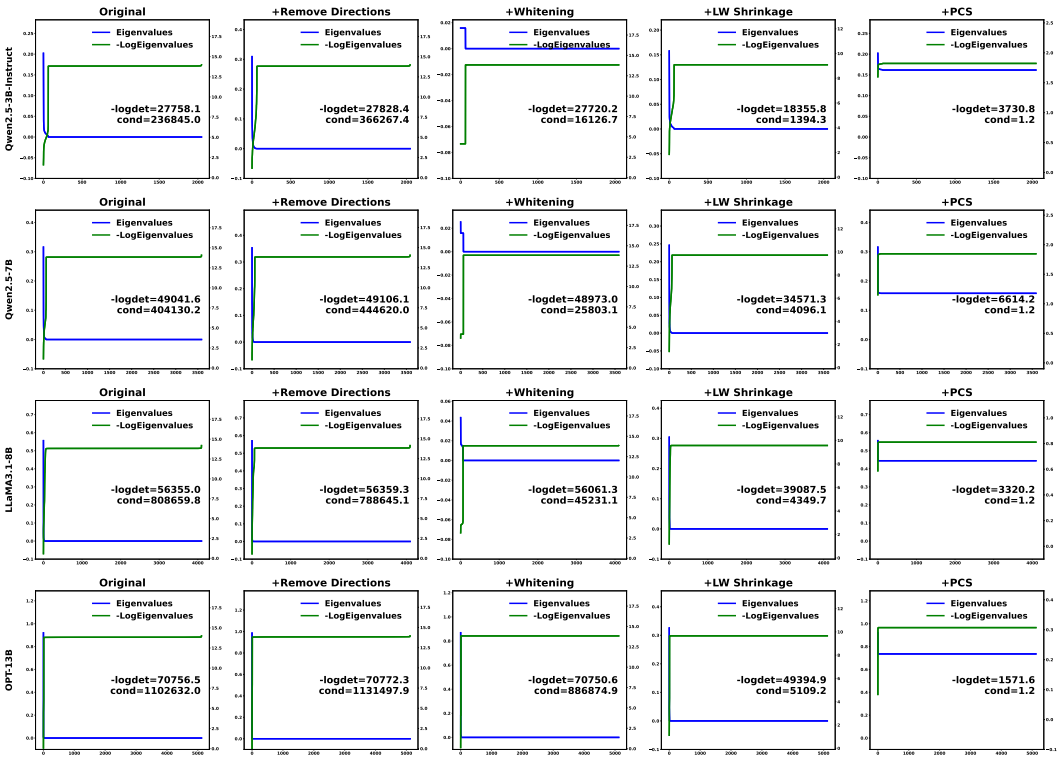


Figure 7: The eigenvalues and their negative logarithmic distributions of different models' representations before and after processing with different anisotropy decoupling methods.

Table 3: The R^2 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.

Model	R^2	**Significance**
LLaMA3.1-8B	0.892	****
LLaMA3.1-8B-Instruct	0.791	***
LLaMA3.2-1B	0.803	****
LLaMA3.2-1B-Instruct	0.788	***
LLaMA3.2-3B	0.891	****
LLaMA3.2-3B-Instruct	0.776	****
OPT-0.125B	0.882	****
OPT-1.3B	0.767	****
OPT-2.7B	0.666	**
OPT-6.7B	0.910	****
OPT-13B	0.830	****
OPT-30B	0.836	****
Qwen2.5-0.5B-Instruct	0.810	****
Qwen2.5-1.5B-Instruct	0.863	***
Qwen2.5-3B-Instruct	0.801	****
Qwen2.5-7B-Instruct	0.800	****
Qwen2.5-14B-Instruct	0.857	****
Qwen2.5-32B-Instruct	0.839	****

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.

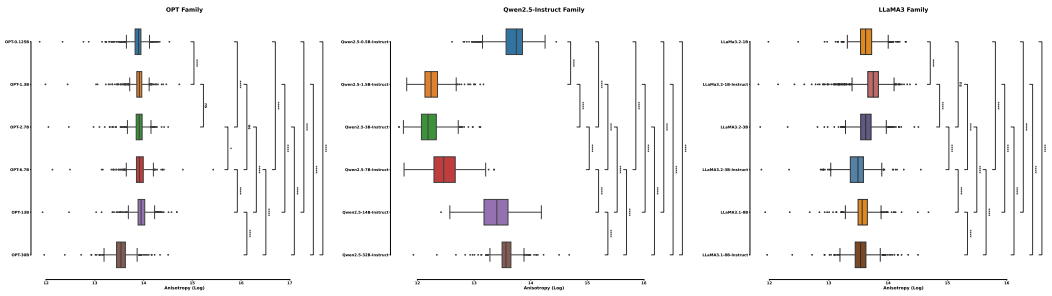


Figure 8: The Mann-Whitney U tests of compression-anisotropy regression fitting between different models, where **** and "ns" denote statistical significance at the 0.01% level and non-significance, respectively.

Figure 8 shows 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.

E Sensitivity Analysis of Evaluation-Framework Hyperparameters

Figure 9 presents the overall pipeline of our intrinsic evaluation framework. Since our goal is to build *task-agnostic* metrics, an important implementation question is whether the framework depends heavily on specific data choices or hyperparameter settings. In this section, we examine two key factors in the framework: the number of projection samples used to elicit representations, and the smoothing coefficient β in principal component smoothing.

For the projection dataset, we intentionally use two representative foundational corpora rather than task-specific benchmarks: a pre-training corpus (Wiki (Foundation, 2025)) and an instruction-tuning dataset (Dolly-15k (Conover et al., 2023)). These two sources cover two common stages in modern LLM development, namely pre-training and instruction tuning. We use a fixed default set of 800 examples to construct the projection data. The purpose here is not to model any particular downstream task, but to activate a broad portion of the model's representation space. In other words, we want the metric to reflect the model's general internal structure, instead of being tied to a narrow task-specific subspace. In this section, we further analyze these design choices from three complementary perspectives: the sensitivity to the smoothing coefficient β, the sensitivity to projection sample size, and the cross-layer behavior of the three metrics.

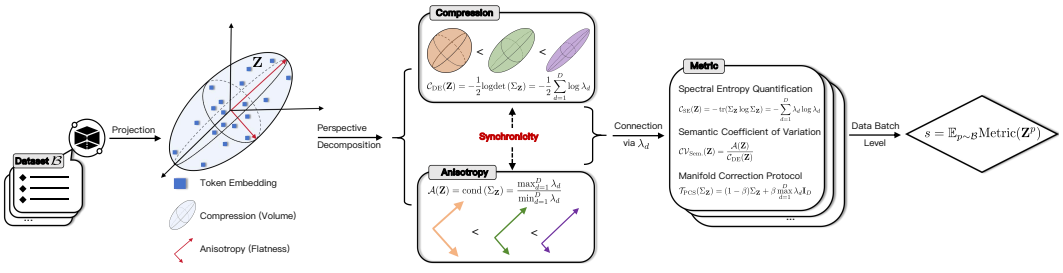


Figure 9: The intrinsic evaluation framework for LLMs based on our proposed metrics.

E.1 Effect of the Smoothing Coefficient β

We first analyze the smoothing coefficient β in principal component smoothing, which controls how strongly the low-eigenvalue tail is corrected in Compression (PCS). Figure 10 shows the correlation between Compression (PCS) and the ground-truth capability scores under different values of β. The non-degenerate interval $\beta \in [0.6, 0.9]$ gives consistently strong correlations of 0.946–0.965, with the default $\beta = 0.9$ reaching 0.965. The endpoints also provide direct algebraic checks. At $\beta = 0$, PCS exactly recovers Compression (DE), and therefore has the same correlation of 0.445. At $\beta = 1$, the transformed covariance is $\lambda_{\max}\mathbf{I}_D$ and PCS reduces to $-\frac{D}{2}\log\lambda_{\max}$; its correlation is 0.779. Thus, the drop at $\beta = 1$ is expected because this boundary case retains only the top eigenvalue, whereas $\beta < 1$ still uses the full spectrum.

Overall, these results show that the framework is not overly fragile to its implementation choices. The projection sample size converges quickly, and Compression (PCS) remains stable across the broad non-degenerate range $\beta \in [0.6, 0.9]$. This gives additional evidence that the proposed framework is practical to use and does not rely on a single smoothing-coefficient setting to obtain strong results.

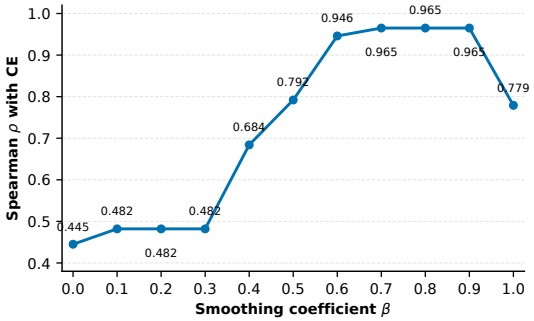


Figure 10: Spearman correlations between Compression (PCS) and comprehensive capability under different smoothing coefficients. At $\beta = 0$, PCS recovers Compression (DE); at $\beta = 1$, it reduces to the top-eigenvalue boundary case $-\frac{D}{2}\log\lambda_{\max}$.

E.2 Effect of Projection Sample Size

We next study the effect of the number of projection samples. As shown in Figure 11, the cumulative expected values of our metrics become stable quickly as the sample size increases. In practice, the curves have already largely converged under the default setting of 800 samples, and adding more samples leads to only minor changes. This result is important for two reasons. First, it shows that the framework is computationally efficient: reliable estimates do not require a very large projection set. Second, it suggests that general foundational corpora are sufficient to produce stable intrinsic measurements, without introducing extra sampling bias from a particular downstream domain. Combined with the strong alignment with external ground truth reported in the main experiments, this supports the view that the intrinsic capability of a model can be estimated robustly from a relatively small amount of general-purpose data.

E.3 Cross-layer Behavior of Our Metrics

To better understand what our metrics capture inside a model, we further examine their layer-wise trajectories. Figure 12 shows the three metrics across layers for LLaMA3.2-3B-Instruct, and extends the same comparison to Qwen2.5-3B-Instruct and OPT-13B using matched trajectories with the correct model depths. Rather than focusing on absolute values alone, this analysis aims to clarify how the three metrics evolve from shallow to deep layers.

A first clear pattern is that Compression (SE) and Compression (PCS) exhibit very similar cross-layer behavior. In all three models, both metrics show an unstable transition in the first few layers, followed by a steady upward trend through the middle and deeper layers. This suggests that the compression-related structure captured by our metrics is increasingly formed as representations become more organized in deeper layers. Put differently, the later layers appear to concentrate the

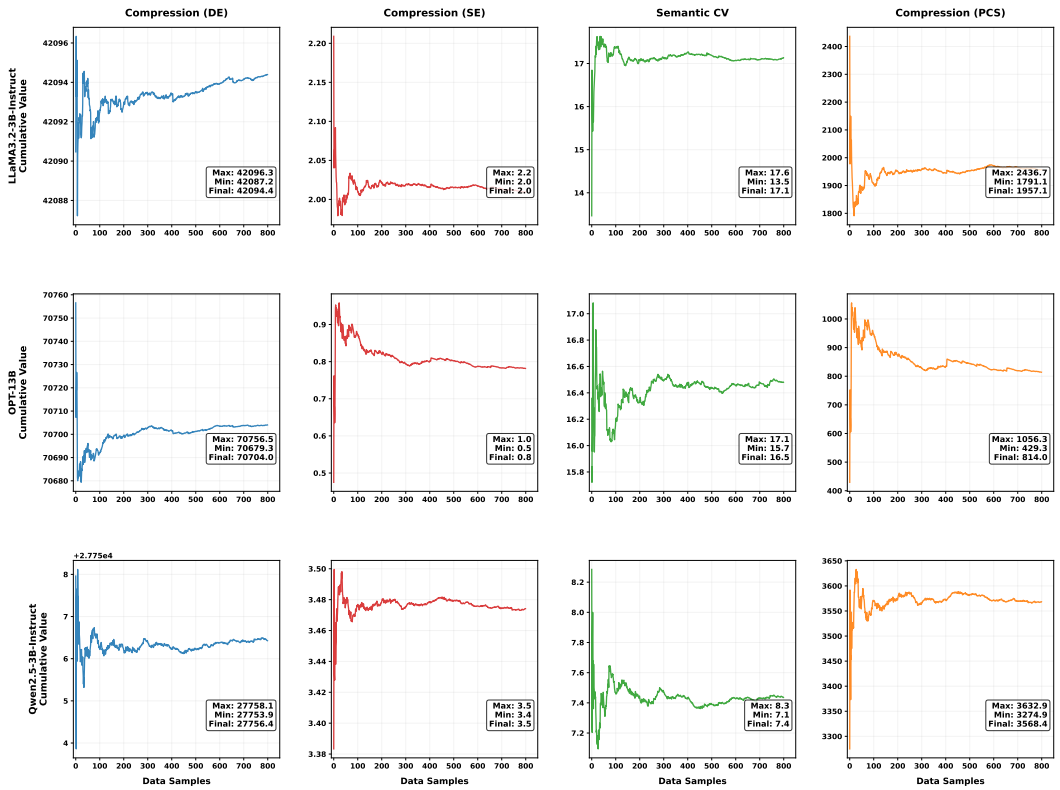


Figure 11: The cumulative expected values of different metrics as the number of samples increases.

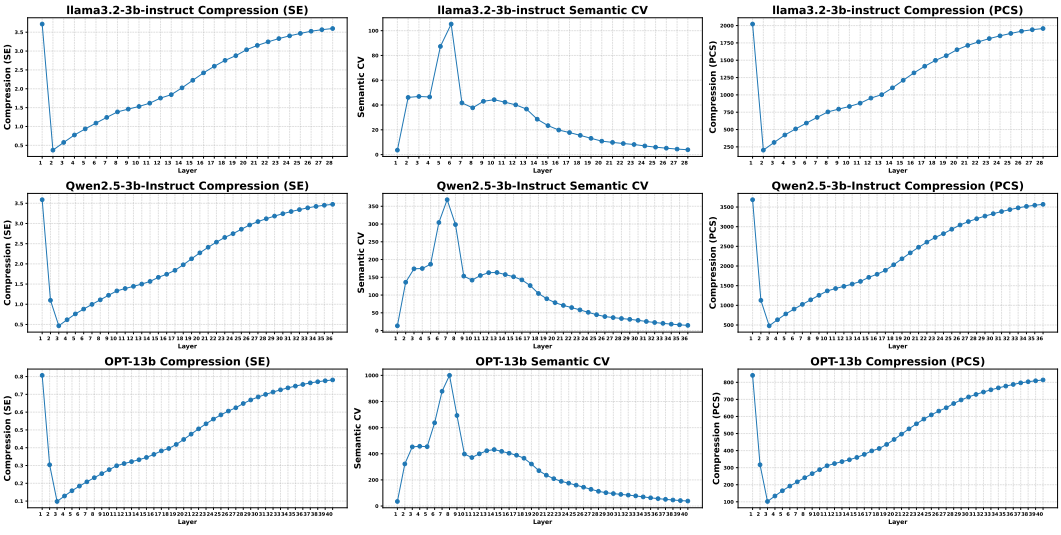


Figure 12: Layer-wise trajectories of Compression (SE), Semantic CV, and Compression (PCS).

informative directions more strongly, which is exactly where compression-based metrics become most discriminative.

Semantic CV shows a different pattern. Instead of continuing to rise toward the final layers, it increases rapidly in the early part of the network, reaches its peak in the early-to-middle layers, and then gradually decreases. This means that the relative strength of anisotropy is most pronounced before the deepest layers are reached. As the representation continues to evolve, the model still becomes more compressed, but the anisotropy-to-compression ratio no longer grows in the same way.

From an intuitive perspective, this suggests that geometric distortion is not simply accumulated layer by layer; rather, it becomes salient at intermediate stages and is then partially absorbed as deeper layers refine the representation.

Another notable observation is that these qualitative trends remain stable across architectures. Although the three models have different depths and different absolute scales, the overall shape of each metric is largely preserved: Compression (SE) rises toward the final layers, Compression (PCS) follows a closely related trajectory, and Semantic CV peaks earlier before decaying. This consistency is encouraging, because it suggests that the proposed metrics are not merely reflecting random layer noise, but are tracking regular structural changes in how LLM representations evolve across depth.

This analysis also provides an intuitive reason for our default use of the last-layer hidden states in the evaluation framework. For the two compression-related metrics, the deepest layers are exactly where the signals become the strongest and most stable. At the same time, the earlier peak of Semantic CV helps explain why geometric distortion is an important complementary factor: it captures a stage of representation organization that is not fully reflected by compression alone.

F Further Analysis of Main Results

F.1 Cross-architecture and Fine-grained Capability Alignment

Table 4: Direction-aligned Spearman correlations within model groups (Qwen2.5-Instruct, OPT, and LLaMA3 families) and across all models (Global), including correlations between the four metrics and both model size (Size) and comprehensive capabilities (Ground truth). Semantic CV is direction-aligned by rank reversal using its reciprocal; the bold-highlighted rows denote our refined metrics.

Metric	Global		Qwen2.5-Instruct Family		OPT Family		LLaMA3 Family	
	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

Figure 4 and Table 4 present the regression equations and Spearman correlations between the original compression metric, our three refined metrics, and comprehensive capabilities as ground truth. The original Compression (DE) shows a strong correlation with model size at the global level (0.935), and this pattern remains nearly perfect within model families (1.000, 1.000, and 0.956). However, its global correlation with comprehensive capability is only 0.445, although it remains relatively high within homogeneous families (0.829, 1.000, and 0.886). This suggests that the original compression metric primarily tracks model scale and is only effective for capability assessment within relatively homogeneous architectural systems.

By contrast, our refined metrics substantially improve cross-architecture capability alignment. Compression (SE) reduces the global size correlation to 0.430 while increasing the global capability correlation to 0.917, indicating that spectral reweighting captures capability differences beyond scale. Semantic CV and Compression (PCS) further strengthen this trend: both maintain high within-family correlations with capability while preserving strong direction-aligned global correlations (0.926 and 0.965, respectively), even though their correlations with size decrease to 0.805 and 0.708. For Semantic CV, the corresponding raw signed capability correlation is -0.926, because lower Semantic CV indicates stronger capability. These results show that our refined metrics are substantially more capability-relevant than the original Compression (DE), and that incorporating geometric distortion extends the "compression as intelligence" perspective beyond model scale alone.

This advantage also persists in fine-grained task evaluations. As shown in Table 2, the original Compression (DE) remains modest and uneven across tasks, with correlations of 0.458 on Factuality, 0.471 on Reasoning, 0.490 on Math, 0.319 on Coding, and 0.521 on Agent. In contrast, our refined metrics achieve uniformly stronger alignment with the ground truth across all evaluation dimensions. The gains are especially pronounced on Factuality and Reasoning, where Compression (SE), Semantic CV, and Compression (PCS) reach 0.895/0.946/0.962 and 0.921/0.926/0.967, respec-

Table 5: Summary statistics of direction-aligned leave-one-model-out Spearman correlations with comprehensive capability, computed from the same 18-model arrays as Figure 4. The full-set value uses all 18 models; the remaining statistics summarize the 18 runs that each retain 17 models. Semantic CV is direction-aligned by rank reversal using its reciprocal.

Metric	Full-set ρ	Mean	Std.	Min	Median	Max	Range
Compression (DE)	0.445	0.444	0.062	0.341	0.447	0.574	0.233
Compression (SE)	0.917	0.916	0.012	0.902	0.912	0.946	0.044
Semantic CV	0.926	0.925	0.012	0.912	0.926	0.963	0.051
Compression (PCS)	**0.965**	**0.964**	**0.007**	**0.958**	**0.962**	**0.983**	**0.025**

tively. Similar improvements are observed on Instruction Following (0.861/0.962/0.955), Knowledge (0.852/0.917/0.923), and Context Learning (0.892/0.916/0.923).

Notably, the refined metrics remain effective on more challenging domains such as Math, Coding, and Agent tasks, where representation-only intrinsic metrics often struggle. On Math and Agent, their correlations consistently remain above 0.82, whereas the original Compression (DE) reaches only 0.490 and 0.521. On Coding, the refined metrics achieve 0.861, 0.875, and 0.875, compared with only 0.319 for the original metric. This suggests that incorporating geometric distortion into compression analysis improves not only overall capability alignment, but also the characterization of domain-specific abilities involving structured reasoning and action-oriented problem solving.

Table 2 also shows that metrics capturing only a single aspect of representations remain systematically weaker. Purely information-based metrics such as Semantic Volume and Compression (DE), as well as purely geometric metrics such as Curvature and anisotropy-based measures, consistently underperform our refined metrics across tasks and on CE. Moreover, simple anisotropy-decoupling variants of Compression (DE), such as Whitening and LW Shrinkage, bring little improvement, while Remove Directions severely degrades performance. In contrast, PCS achieves the strongest overall results, with the highest CE correlation of 0.965. Overall, these findings support our central claim that jointly modeling information compression and geometric distortion yields a more faithful intrinsic proxy for LLM capability.

F.2 Robustness under Leave-One-Model-Out Analysis

To further examine whether the observed capability alignment is robust to the composition of the evaluated model set, we conduct a leave-one-model-out (LOO) analysis using exactly the same 18-model arrays as Figure 4. For each metric, we remove one model, recompute the Spearman correlation with the comprehensive capability score on the remaining 17 models, and obtain 18 LOO correlations. Table 5 summarizes these distributions, while Table 6 reports every held-out case.

Compression (DE) has a full-set correlation of 0.445 and a mean LOO correlation of 0.444. Its LOO values range from 0.341 to 0.574 with a standard deviation of 0.062, so its capability alignment is both substantially weaker and more sensitive to model composition than those of the refined metrics.

Compression (SE) has a full-set correlation of 0.917 and a mean LOO correlation of 0.916; its 18 LOO values lie between 0.902 and 0.946 with a standard deviation of 0.012. Direction-aligned Semantic CV similarly has a full-set correlation of 0.926 and a mean LOO correlation of 0.925, with values between 0.912 and 0.963 and a standard deviation of 0.012. Compression (PCS) is strongest and most concentrated: its full-set and mean LOO correlations are 0.965 and 0.964, respectively, and all LOO values fall between 0.958 and 0.983 with a standard deviation of 0.007. Thus, the refined metrics remain well above DE regardless of which single model is removed.

This robustness pattern is consistent with the task-level results in Table 2. We interpret LOO as a sensitivity analysis, not as 18 independent experiments: any two LOO subsets share 16 of their 17 models. We therefore do not attach independent-sample significance claims to these values. The direct evidence is the complete set of recomputed correlations in Table 6, which makes both the effect size and its variation under model removal transparent.

Table 6: Detailed direction-aligned leave-one-model-out Spearman correlations computed from the same data as Figure 4. Column $-i$ reports the correlation after removing the ith model in the released 18-model data array; Semantic CV is direction-aligned by rank reversal using its reciprocal.

Metric	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16	-17	-18
Compression (DE)	0.341	0.547	0.574	0.436	0.419	0.341	0.341	0.417	0.417	0.488	0.507	0.468	0.456	0.456	0.439	0.412	0.466	0.466
Compression (SE)	0.931	0.912	0.907	0.912	0.907	0.902	0.902	0.912	0.919	0.931	0.936	0.917	0.909	0.914	0.917	0.946	0.907	0.907
Semantic CV	0.912	0.926	0.926	0.912	0.926	0.912	0.912	0.912	0.922	0.929	0.929	0.963	0.926	0.941	0.924	0.922	0.926	0.926
Compression (PCS)	0.958	0.968	0.958	0.958	0.958	0.958	0.958	0.958	0.958	0.966	0.968	0.983	0.968	0.968	0.958	0.973	0.968	0.968

G Related Lemma and Proposition

Lemma G.1 (Differential Entropy of Gaussian Vectors)**.** *For $\mathbf{x} \in \mathbb{R}^D$ that follows a multivariate normal distribution $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, The differential entropy of $\mathbf{X}$ is given by:*

$$\mathcal{H}_{DE}(\mathbf{x}) = \frac{1}{2}\log(\det(\mathbf{\Sigma})) + C$$

where $C = \frac{D}{2}(\log(2\pi) + 1)$ is a constant term.

Proof. For a continuous random variable $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^D$, the differential entropy is defined as:

$$\mathcal{H}_{\text{DE}}(\mathbf{x}) = -\int f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x}$$

The probability density function (PDF) is given by:

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

First, we compute the logarithm of $f(\mathbf{x})$:

$$\begin{aligned}\log f(\mathbf{x}) &= \log\left(\frac{1}{\sqrt{(2\pi)^D \det(\mathbf{\Sigma})}}\right) - \frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) \\ &= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{\Sigma})) \\ &\quad - \frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\end{aligned}$$

Substituting $\log f(\mathbf{x})$ into the definition of differential entropy, we obtain:

$$\mathcal{H}_{\text{DE}}(\mathbf{x}) = -\int f(\mathbf{x})(-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log(\det(\Sigma)) - \frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu))\, d\mathbf{x}$$

Since $f(\mathbf{x})$ is normalized (its integral over the entire space is 1), the constant terms can be pulled out of the integral:

$$\mathcal{H}_{\text{DE}}(\mathbf{x}) = \frac{D}{2}\log(2\pi) + \frac{1}{2}\log(\det(\Sigma)) + \frac{1}{2}\int f(\mathbf{x})(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\, d\mathbf{x}$$

The integral term represents the expectation of the quadratic form $\mathbb{E}[(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)]$. Using the trace property of expectations, we have:

$$\int f(\mathbf{x})(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\, d\mathbf{x} = \text{tr}(\Sigma^{-1}\Sigma) = D$$

where n is the dimension of the variable. Thus, the entropy simplifies to:

$$\begin{aligned}\mathcal{H}_{\text{DE}}(\mathbf{x}) &= \frac{D}{2}\log(2\pi) + \frac{1}{2}\log(\det(\Sigma)) + \frac{D}{2} \\ &= \frac{1}{2}\log(\det(\Sigma)) + \frac{D}{2}(\log(2\pi)+1) \\ &= \frac{1}{2}\log(\det(\Sigma)) + C\end{aligned}$$

where $C = \frac{D}{2}(\log(2\pi) + 1)$ is a constant term. □

Proposition G.2 (Spectral Stabilization of PCS)**.** *Let $\Sigma_{\mathbf{Z}}$ be positive definite with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D > 0$, and define*

$$\mathcal{T}_{PCS}(\Sigma_{\mathbf{Z}}) = (1-\beta)\Sigma_{\mathbf{Z}} + \beta\lambda_1 \mathbf{I}_D, \qquad \beta \in (0, 1].$$

Then the eigenvalues of $\mathcal{T}_{PCS}(\Sigma_{\mathbf{Z}})$ are

$$\tilde{\lambda}_d = (1-\beta)\lambda_d + \beta\lambda_1, \qquad d = 1, \ldots, D.$$

Consequently,

$$\mathrm{cond}\big(\mathcal{T}_{PCS}(\Sigma_{\mathbf{Z}})\big) = \frac{\tilde{\lambda}_1}{\tilde{\lambda}_D} = \frac{\lambda_1}{(1-\beta)\lambda_D + \beta\lambda_1} \leq \min\left\{\mathrm{cond}(\Sigma_{\mathbf{Z}}),\ \frac{1}{\beta}\right\}.$$

Moreover, for all $d \geq 2$,

$$\tilde{\lambda}_d \geq \beta\lambda_1,$$

so the tail eigenvalues are uniformly lifted away from zero. Therefore, in sparse-spectrum settings where small eigenvalues disproportionately dominate the log-det term, PCS suppresses the instability induced by near-zero spectral directions and makes the resulting compression score less sensitive to anisotropy.