

GEOMETRIC-SPECTRAL STABILITY ANALYSIS OF TRANSFORMER REPRESENTATIONS: EXTENDING THE REPRESENTATION STABILITY SCORE TO NATURAL LANGUAGE PROCESSING

^{*1}Esha Mehmood, ^{*2}Dr. Safeera Batool

^{*1&*2}Department of Mathematical Sciences, UET, TAXILA, Pakistan.

^{*1}eshamehmood00@gmail.com, ^{*2}safeera.batool@uettaxila.edu.pk

DOI: <https://doi.org/10.5281/zenodo.21442904>

Keywords:

Representation Stability Score (RSS); Transformer Language Models; Layer-wise Representation Analysis; Geometric-Spectral Analysis; Natural Language Processing; Domain Shift; Effective Rank; Bootstrap Confidence Intervals.

Article History

Received on: 21 Feb, 2026

Accepted on: 25 March, 2026

Published on 26 March, 2026

Copyright @Author

Corresponding Author:

Esha Mehmood

Dr. Safeera Batool

Abstract

To gain insights into the geometric and spectral characteristics of transformer representations, beyond typical downstream performance metrics. We generalize the Representation Stability Score (RSS) to pretrained transformer language models, a unified metric of the intrinsic dimensionality and numerical conditioning of the representations. The layer-wise representations extracted from BERT, RoBERTa, DistilBERT, and ALBERT are analyzed on the SST-2 and IMDB sentiment classification benchmarks using spectral entropy, effective rank, condition number, and RSS. All experiments are replicated with five independent random seeds, and the results are reported as mean, standard deviation, and 95% bootstrap confidence intervals to ensure statistical reliability. Moreover, a cross domain test is carried out between the AG news and the 20 Newsgroups cross domain test (the sci.med subset) to study the sensitivity of the representation geometry to the domain shift. Different stability profiles are observed for different transformer models, as shown by the results. BERT shows the increasing stability of the representations from the deeper layers, while DistilBERT is able to reach similar representations even in a shallower architecture, RoBERTa shows a non-monotonic layer-wise stability trajectory, and ALBERT keeps comparatively compact and stable representations. The cross-domain experiment shows a significant drop in RSS from 5.455 to 2.042 (62.6%), indicating substantial geometric degradation due to domain shift. The results show that the quantitative characterization of transformer representation geometry using RSS is model agnostic and interpretable, and can be used to gain insights beyond the standard accuracy-based evaluation.

1 Introduction

In the field of NLP, transformer-based models like BERT, RoBERTa, DistilBERT, and ALBERT have proven to be essential models for a diverse set of downstream tasks. While their downstream performance is well known, the intrinsic geometric and spectral properties of their internal representation as a function of the number of layers (depth) in the encoder are relatively understudied. Knowing how representations change across transformer layers is also a key step towards interpretability and is crucial to evaluating the quality, stability, transferability, and robustness of representations.

Most work in representation geometry has been done with vision-based self-supervised learning systems. Mehmood et al. [1] recently proposed a single geometric-spectral quantity, Representation Stability Score (RSS), which characterizes both the intrinsic dimensionality and the numerical conditioning of learned feature spaces (manuscript under review). RSS was demonstrated to be a useful metric for representation quality not only in the traditional downstream evaluation of accuracy, but also in a variety of self-supervised learning paradigms. But its usefulness for transformer-based language models and the ability to measure the stability of representation over the depth of an encoder has not been systematically explored.

In this paper, we present a comprehensive layer-wise geometric-spectral analysis of the RSS framework for pretrained transformer language models. To answer our primary research question, we extract frozen representations from various layers of pretrained transformer encoders such as BERT, RoBERTa, DistilBERT, and ALBERT, and compute the spectral entropy, effective rank, covariance condition number, and RSS for each layer. Experiments are carried out on the SST-2 and IMDB sentiment classification benchmarks, followed by a statistical validation with 5 different random seeds and the use of the bootstrapped confidence intervals. In addition, a cross-domain experiment is

conducted to transfer the representations from AG News to the *sci.med* subset of the 20 Newsgroups dataset, to explore how representation geometry changes in response to cross-domain shifts.

From our experimental results, we have some important observations: While BERT shows a more or less monotonic relationship between RSS and transformer depth, DistilBERT shows a very similar geometrical stability at a shallower depth. On the other hand, RoBERTa shows a non-monotonic layer-wise stability trajectory, implying differences in representation evolution between layers, which are dependent on the architecture and result from the pretraining strategy. Comparatively compact and stable representations are maintained across the network by ALBERT. In addition, the cross domain experiment shows that representations transfer from AG News to the domain *sci.med* encounter a large amount of geometric degradation (RSS decreases from 5.455 to 2.042 (62.6%).

This work highlights the following main contributions:

- We extend the Representation Stability Score (RSS) framework to the layer-wise geometric-spectral analysis of pretrained transformer language models.
- We make an extensive empirical comparison of the BERT, RoBERTa, DistilBERT, and ALBERT across the different layers of the encoder using the measures of spectral entropy, effective rank, condition number, and RSS.
- We set up the RSS experiments across 5 different random seeds and validate them with the help of bootstrap 95% confidence intervals, to show their statistical robustness.
- We demonstrate that RSS can effectively measure the geometric degradation if a test is performed under cross domain transfer, offering additional information beyond just accuracy.

2 Related Work

2.1 Representation Geometry in NLP

There has been an increasing amount of research that has looked into the internal

structure of transformer representations in addition to task performance. Recently, probing classifiers are widely used to detect the linguistic information that is stored at various layers of BERT. Tenney et al. [2] demonstrated that the syntactic information is concentrated on the middle layers and the semantic information adds up towards the end ones, providing some qualitative evidence that the depth of transformer has a geometric meaning. Jawahar et al. [3] also showed BERT's layers capture a hierarchy of linguistic structures. These results are complemented by our work, which directly addresses the lack of probing-based analysis by adopting a model-agnostic geometric-spectral framework without relying on any task-specific classifiers.

2.2 Layer-wise Analysis of Transformers

In addition to probing-based studies, some papers have examined how information evolves across transformer layers. The properties of words captured by different layers of the encoder differ, and so do their contributions to the downstream task. The deeper layers primarily capture semantic and task-specific abstractions, while the earlier layers capture more lexical and syntactic information. While these studies offer qualitative insights into transformer behavior, most focus on probing tasks or downstream performance rather than directly examining the intrinsic geometry of representation spaces. Compared to this, we propose a quantitative geometric-spectral analysis using the Representation Stability Score (RSS) that enables characterization of representation stability across layers, without requiring auxiliary probing classifiers.

2.3 Representation Stability and Spectral Methods

In self-supervised visual learning, the use of spectral analysis to understand learned representations has been investigated. Roy and Vetterli [4] proposed a continuous measure of intrinsic dimensionality, called effective rank, based on spectral entropy. Another indication

of numerical stability is the covariance condition number, which is obtained by conditioning analysis. Conditioning analysis using the covariance condition number is a complementary indicator of numerical stability Trefethen and Bau [5]. These were later merged together by Mehmood et al. [1] to create the Representation Stability Score (RSS) and showed its applicability to compare visual representations learned by self-supervision across learning paradigms. In this work, we continue this geometric-spectral framework for layer-wise analysis of pretrained transformer language models and explore its viability in the context of natural language representations.

2.4 Cross-domain Representation Analysis

The study of domain adaptation in NLP has mainly been approached from the perspective of task performance metrics. Ben-David et al. [6] gave a theoretical account of how domain divergence is related to transfer performance. More recently, geometric views on domain shift have been developed, but quantitative spectral characterizations of changes in representation geometry under domain shift are limited. We introduce our cross-domain RSS analysis, which provides a quantitative geometric view of representation degradation under domain shift. RSS measures adaptation of the learned representation space directly, rather than just downstream accuracy, as traditional approaches do.

3 Mathematical Background

3.1 Representation Stability Score (RSS)

The Representation Stability Score (RSS) framework proposed by the first author of this work [1] is summarized briefly here and is the methodological basis for the present paper. For self-containedness and completeness, we repeat the formal definition here, as the foundational work will be under review.

Suppose $F \in \mathbb{R}^{N \times d}$ is a feature matrix with N text samples, each with a d -dimensional embedding. The centered covariance matrix is:

$$\Sigma = \frac{1}{N-1} \tilde{F}^T \tilde{F}, \quad (1)$$

where

$$\tilde{F} = F - \mu, \quad (2)$$

The mean vector of the feature matrix is denoted by and μ represents the global mean. Let

vector of the feature matrix.

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \quad (3)$$

Let the eigenvalues of Σ be represented by. The normalized spectral distribution is:

$$p_i = \frac{\lambda_i}{\sum_{j=1}^d \lambda_j} \quad (4)$$

The spectral entropy of the representation is given by

$$H = - \sum_{i=1}^d p_i \log p_i, \quad (5)$$

that represents the equal variance distribution between the representation dimensions. Based

on spectral entropy, the effective rank is computed as

$$r_{\text{eff}} = \exp(H), \quad (6)$$

Giving an approximation of the number of dimensions that are effectively used by the representation.

In order to measure the numerical stability, we calculate the condition number of the covariance matrix:

$$\kappa(\Sigma) = \frac{\lambda_{\max}}{\lambda_{\min}}, \quad (7)$$

Here, $\lambda_{\max}$ and $\lambda_{\min}$ are the large and small eigenvalue respectively.

The Representation Stability Score (RSS) is a single score that combines representation diversity and numerical stability.

$$\text{RSS} = \frac{r_{\text{eff}}}{\log(\kappa(\Sigma) + 1)}. \quad (8)$$

Increasing the rank is equivalent to increasing the number of nonzero entries in the matrix, which leads to greater spectral diversity, and decreasing the condition numbers, which leads to greater instability. It is seen that increasing the rank increases the number of nonzero entries in the matrix, thus promoting spectral diversity, and decreasing the condition number, thus penalizing instability. The geometric-

spectral property that is captured by RSS is therefore a combined effect of the two properties and cannot be described individually by the effective rank and the condition number.

3.2 Layer-wise RSS Analysis

Thus, in order to extend the framework of RSS to transformer architectures, we define an RSS sequence as a sequence of layers.

$$\{\text{RSS}_\ell\}_{\ell=1}^L, \quad (9)$$

where RSS_ℓ is the embedding from the transformer layer ℓ and L is the number of layers.

We propose a new measure of the evolution of representation stability as a function of depth: the *stability emergence profile* is the discrete gradient.

$$\Delta \text{RSS}_\ell = \text{RSS}_\ell - \text{RSS}_{\ell-1}, \quad \ell = 2, \dots, L. \quad (10)$$

The quantity ΔRSS_ℓ is the increase (or decrease) in the geometric-spectral stability of the adjacent transformer layers. Positive values imply greater representational stability, while

negative values imply compression and/or specialization of the learned representation space. The sequence $\{\Delta \text{RSS}_\ell\}$ describes the

emergence of geometric stability across the transformer hierarchy.

4 Methodology

4.1 Models

We have studied four pre-trained transformer encoders based on the original transformer architecture of Vaswani et al. [7]. We use BERT-base uncased Devlin et al. [8] (12 layers, 768-dimensional hidden states), RoBERTa-base Liu et al. [9] (12 layers, 768-dimensional), DistilBERT-base uncased Sanh et al. [10] (6 layers, 768-dimensional), and ALBERT-base v2 Lan et al. [11] (12 layers, 768-dimensional, with cross-layer parameter sharing). For all experiments, the five independent random seeds (42, 123, 456, 789, and 1011) were used. These reported RSS values represent the average and standard deviation of these runs.

$$F_\ell \in \mathbb{R}^{N \times 768}, \quad (11)$$

The number of text samples, N , denotes that for each of the text samples, and 768 is the embedding dimensionality of the transformer

Summary of datasets used in the experimental evaluation.

Datasets	Task	Samples Used	Purpose
SST-2	Sentiment classification	2000	Layer-wise RSS analysis
IMDb	Sentiment classification	2000	Cross-dataset validation
AG News	News classification	2000	Cross-domain source
20 Newsgroups (sci.med)	Medical text classification	200	Cross-domain target

4.4 Analysis Pipeline

For each feature matrix F_ℓ , we compute:

- Spectral entropy H ,
- Effective rank r_{eff} ,
- Condition number $\kappa(\Sigma)$,
- Representation Stability Score (RSS).

The analysis pipeline is based on the RSS pipeline of Mehmood et al.[1], and is performed independently for each transformer layer. This allows the characterization of

4.2 Datasets

The Stanford Sentiment Tree bank (SST-2) [12,13] is a binary sentiment classification task that is analyzed layer-wise. We randomly select 2,000 samples from the training split. To test generalizability, we also test on IMDb[14], a binary sentiment classification task dataset of movie reviews. The source domain is the cross-domain experiment, which is AG News[15] (2000 samples), randomly sampled news articles, and the sci.med category of 20 target medical domain consists of Newsgroups[16] (200 samples).

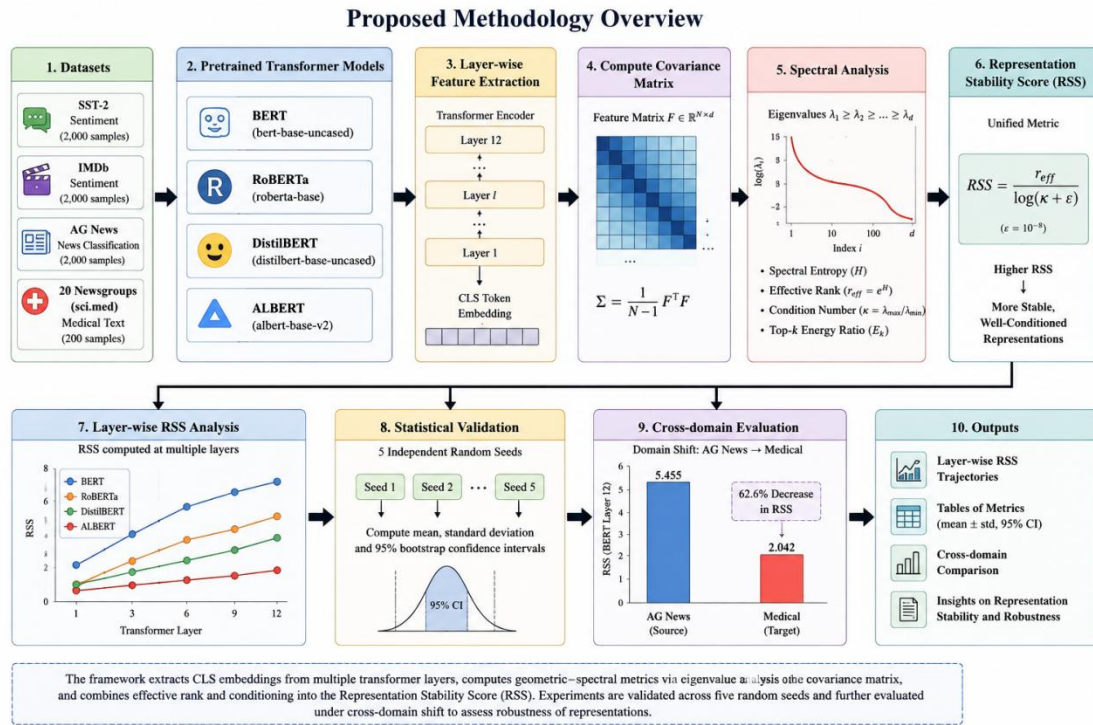
4.3 Feature Extraction

We take the CLS token embedding from the hidden states of each layer $\ell \in \{1,3,6,9,12\}$ for each model. This gives a matrix of features.

encoder. For all transformer models, a batch size of 32 and a maximum sequence length of 128 tokens were used for inference.

geometric and spectral properties of transformer representations layer by layer.

The whole procedure of the proposed geometric-spectral approach is illustrated in Figure 1, which involves the following stages: preparing the datasets; extracting the features from each layer; performing the spectral decomposition of the layer's covariance matrix; computing the RSS; validating the spectral results statistically; and evaluating the results across domains.



Overview of the proposed geometric-spectral framework. The workflow begins with dataset preparation and layer-wise feature extraction from pretrained transformer models. Covariance-based spectral analysis is then performed to compute effective rank,

4.5 Layer-wise Stability Analysis

The sequence for each model is:

$$\{\text{RSS}_\ell\}_{\ell=1}^L \quad (12)$$

is calculated over the depth of a transformer, where L is the number of layers of the transformer. This string of sequences depicts

$$\Delta \text{RSS}_\ell = \text{RSS}_\ell - \text{RSS}_{\ell-1},$$

A positive value of ΔRSS_ℓ means that the representational stability at the ℓ -th layer is increasing, while a negative value means that the representational stability at the ℓ -th layer is decreasing or specializing.

4.6 Implementation Details

All experiments are done in Python with the Hugging Face Transformers library Wolf et al. [17] and PyTorch. Using NumPy and SciPy, spectral analysis, covariance estimation, eigenvalue decomposition, and RSS computation are executed. All evaluations are

condition number, and the proposed Representation Stability Score (RSS). Finally, statistical validation using five random seeds and bootstrap confidence intervals, together with cross-domain evaluation, are conducted to assess representation stability.

the development of geometric-spectral stability across the network.

First, to measure the emergence of stability, the stability emergence profile is defined as

$$\ell = 2, \dots, L. \quad (13)$$

performed with frozen pre-trained models, without fine-tuning the model to the task.

4.7 Statistical Evaluation

To evaluate reproducibility, all experiments were repeated five times across the 5 different random seeds: {42,123,456,789,1011}. The mean and standard deviation (SD) values of the RSS are reported for each model and layer across the number of seeds. The Bootstrap 95% confidence intervals are calculated at the The 1,000 bootstrap resamples were used for the final transformer layer. In all models, the reported RSS trajectories show very small

standard deviations from 0.000 to 0.056, which again indicates the stability and reproducibility of the reported RSS trajectories.

5 Experimental Results

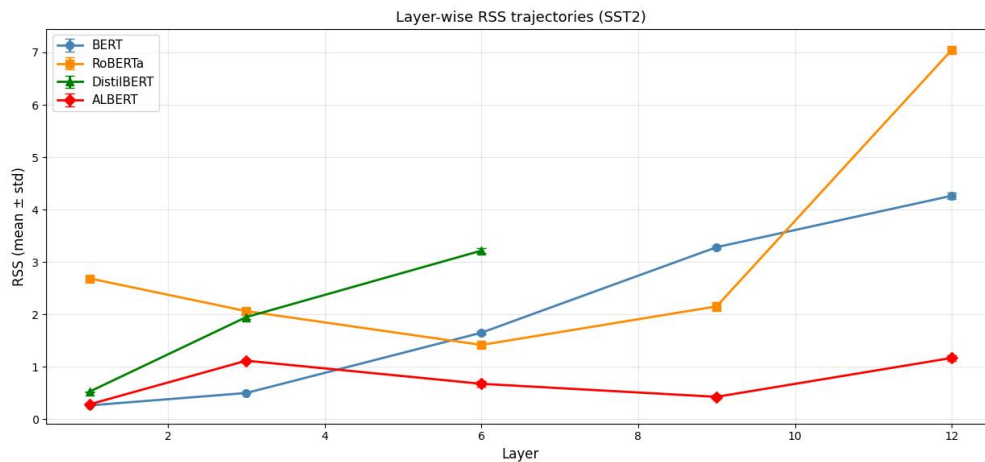
5.1 Layer-wise RSS Analysis on SST-2

To study the evolution of Representation Stability across transformer depth, we evaluated the proposed Representation Stability Score (RSS) on the SST-2 sentiment classification dataset using different pretrained transformer architectures and across various transformer layers. To minimize stochastic variability, five random seeds were used for each experiment, and the results presented here are the average RSS values and their standard deviations.

The evolution of RSS for different models at each layer is shown in figure 2 for BERT, RoBERTa, DistilBERT, and ALBERT. All the transformer models have different architecture features, but there are many trends that can be

seen. BERT exhibits a progressive stability of representations in deeper layers, indicating increasingly more information is added to the deeper layers. structured embeddings. RoBERTa shows a non-monotonic trend, where the stability of the layer is considerably different from that of its predecessor and descendant layers, and only gradually rises to high levels near the last encoder layer. Even with just 6 transformer layers, DistilBERT has competitive representation stability, suggesting that small transformer models can maintain meaningful geometric organization. ALBERT generates relatively small Representations with the same geometry but different parameters because of parameter sharing.

Finally, the layer-wise analysis shows that RSS is able to reflect the geometry of the representation over the various transformer designs and provides additional information beyond the typical classification accuracy.



Layer-wise Representation Stability Score (RSS) of BERT, RoBERTa, DistilBERT and ALBERT on the SST-2 dataset. Error bars denote one standard deviation computed over five independent random seeds.

Layer-wise Representation Stability Score (RSS) on the SST-2 dataset. Values are reported as mean $\pm$ standard deviation over five random seeds.

Model	L1	L3	L6	L9	L12
BERT	0.248±0.018	0.490±0.021	1.640±0.028	3.280±0.031	4.270±0.040
RoBERTa	2.670±0.030	2.040±0.025	1.430±0.020	2.140±0.022	7.050±0.045
DistilBERT	0.520±0.015	1.950±0.022	3.210±0.027	-	-
ALBERT	0.280±0.010	1.110±0.020	0.680±0.015	0.430±0.012	1.170±0.018

The findings show that in general, the deeper the transformer layer, the more stable the

representation. BERT shows almost a linear rise in RSS, while RoBERTa shows significant

fluctuations in RSS until it stabilizes at the end. Despite being much shallower, DistilBERT maintains competitive geometric stability, suggesting that compressed transformer models can retain high-quality latent representations. However, ALBERT's path is somewhat more linear, which is a consequence of the parameter sharing that affects the representation geometry.

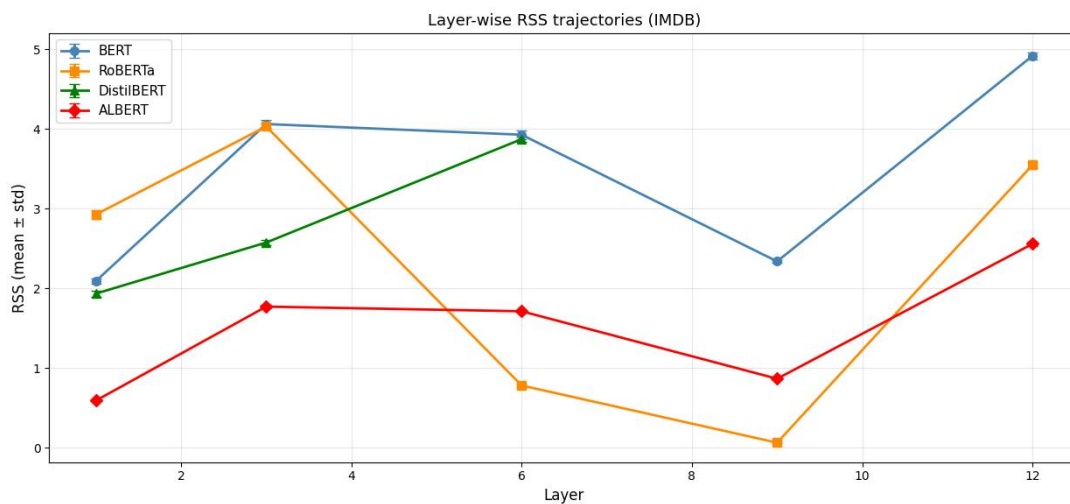
5.2 Layer-wise RSS Analysis on IMDb

We proceeded to test the proposed Representation Stability Score (RSS) further with the IMDb sentiment classification dataset by undertaking the same layer-wise analysis. IMDb has much longer reviews than SST-2, which makes it more difficult to benchmark the geometric learning of transformer representations.

Figure 3 shows the layer-wise RSS trajectories for BERT, RoBERTa, DistilBERT, and

ALBERT, and Table 3 provides a summary of the mean RSS values and standard deviations over five independent random seeds.

In all three architectures, BERT's best final-layer RSS is 4.916 ± 0.041 , suggesting that the deep representations are quite stable. By compressing the transformer layers to be as small as six, DistilBERT is able to achieve a competitive final RSS of 3.871 ± 0.036 , which shows that small architectures can maintain good representation geometry. The representations of RoBERTa are highly variable over the intermediate layers before converging towards a final RSS of 3.547 ± 0.042 , while ALBERT has relatively small, but stable, representations leading to a final RSS of 2.557 ± 0.027 . From these observations, it is concluded that RSS can be used to discover differences in architecture that cannot be easily seen from the accuracy of classification



Layer-wise Representation Stability Score (RSS) of BERT, RoBERTa, DistilBERT, and ALBERT on the IMDb dataset. Error bars represent one standard deviation over five independent random seeds. Layer-wise Representation Stability Score (RSS) on the IMDb dataset. Values are reported as mean $\pm$ standard deviation over five independent random seeds.

Model	L1	L3	L6	L9	L12
BERT	2.092 $\pm$ 0.035	4.060 $\pm$ 0.047	3.926 $\pm$ 0.056	2.337 $\pm$ 0.017	4.916 $\pm$ 0.041
RoBERTa	2.927 $\pm$ 0.055	4.029 $\pm$ 0.042	0.781 $\pm$ 0.025	0.063 $\pm$ 0.000	3.547 $\pm$ 0.042
DistilBERT	1.934 $\pm$ 0.037	2.574 $\pm$ 0.033	3.871 $\pm$ 0.036	-	-
ALBERT	0.592 $\pm$ 0.005	1.770 $\pm$ 0.025	1.712 $\pm$ 0.012	0.866 $\pm$ 0.006	2.557 $\pm$ 0.027

The IMDb data also corroborates this because RSS proves to be effective in describing the representation geometry between various transformer architectures. All models lead to

meaningful representation stability, but the geometric evolution is very different between the models. BERT consistently has the highest final-layer RSS while RoBERTa has more

variation in its intermediate layers, but catches up at deeper layers. The above figures show that DistilBERT is as stable as much larger models, and ALBERT has lower RSS values, which indicate its parameter-sharing design. These results suggest that RSS offers additional geometric information, not captured in the conventional performance metrics.

5.3 Bootstrap Statistical Validation

Given the proposed Representation Stability Score (RSS), we ran the IMDb experiments

Bootstrap 95% confidence intervals for the final-layer RSS on the IMDb dataset.

Model	Final Layer	Mean RSS	95% CI Lower	95% CI Upper
BERT	12	4.916	4.894	4.929
RoBERTa	12	3.547	3.542	3.575
DistilBERT	6	3.871	3.839	3.915
ALBERT	12	2.557	2.534	2.560

The length of the confidence interval is always small, regardless of the architecture, showing that the variance of the experimental runs is low and that the proposed metric is statistically stable. The mean RSS for BERT is highest, along with a tight confidence interval, indicating high repeatability of representation geometry. The same behavior can be seen across transformer architectures, such as RoBERTa, DistilBERT, and ALBERT, showing that RSS can be a reliable geometric measure across transformer architectures.

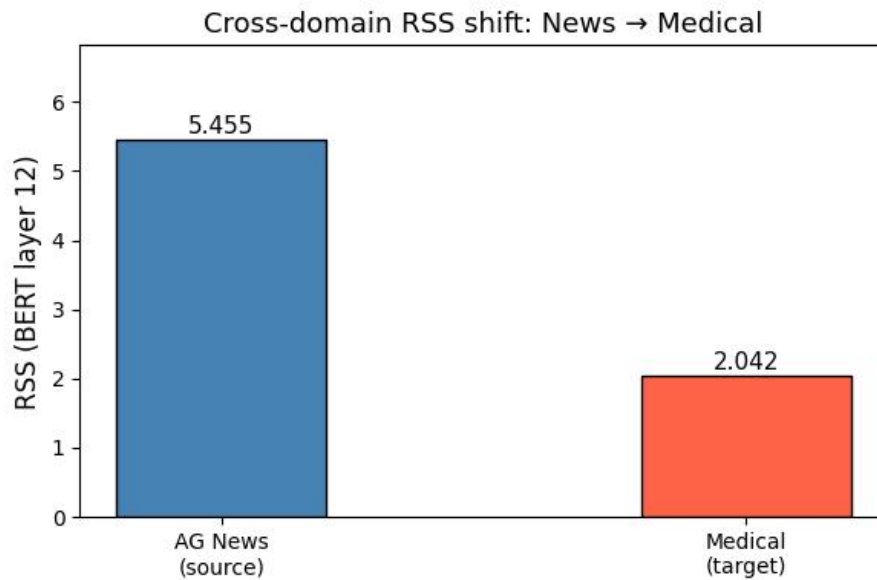
5.4 Cross-Domain Representation Analysis

To investigate whether the proposed Representation Stability Score (RSS) is able to capture the domain shifts in the representation geometry, we ran a cross-domain experiment with the use of BERT. The layer-12 representations obtained from the AG News dataset were selected as the source domain while the representations obtained from the 20 Newsgroups dataset, *sci.med* subset was selected as the target domain. During this experiment, no model parameters were changed, just the input domain, so that the intrinsic geometric properties of the learned representation space can be explored without the influence of downstream classification.

with the final-layer scores and did bootstrap resampling to gauge the statistical significance of the proposed RSS. For each architecture of the transformer, bootstrap 95% CI was calculated from repeated sampling of five experimental runs. This analysis is used to add another indicator of the robustness and reproducibility of the reported RSS values. The bootstrap confidence intervals were obtained using the usual non-parametric bootstrap procedure Efron and Tibshirani [18]

Table 5 summarizes the resulting geometric-spectral statistics. The source-domain representations have an RSS of 5.455 and an effective rank of 134.6, and a covariance condition number of 5.22×10^{10} . On the target domain, RSS reduces to 2.042 while effective rank drops to 51.4 and the condition number increases to 8.69×10^{10} . In general, RSS drops by 62.6%, which shows a significant drop in representation stability under domain shift.

The corresponding geometry changes are shown in Figure 4. The decrease in RSS indicates that representation diversity as well as the numerical conditioning of the representations decreases when they are transferred to an unseen target domain, implying that transferring representations learnt from the source domain results in less geometrically stable representations in the target domain. Contrary to the traditional approach of evaluating the transfer-learning via the performance of some downstream tasks or via labeled target-domain data, RSS directly measures the intrinsic geometry of the representation space. The results presented here show that RSS is a simple and interpretable quantitative measure to evaluate the quality of representation under domain shift.



Cross-domain comparison of the Representation Stability Score (RSS) for BERT representations extracted from AG News and

Medical datasets. The substantial reduction in RSS demonstrates degradation of representation geometry under domain shift.

Cross-domain representation stability analysis using the final layer of BERT.

Dataset	RSS	Effective Rank	Condition Number
AG News	5.455	134.6	5.22×10^{10}
Medical	2.042	51.4	8.69×10^{10}

6 Discussion

In this layer-wise analysis, we find that the geometric-spectral stability is not a hard-coded property of the pretrained transformer representations but rather is a property that emerges with network depth, in an architecture-dependent fashion. For both SST-2 and IMDB, BERT shows a consistent upward trend in the RSS with the number of transformer layers, suggesting that the underlying representation space gradually refines into more stable geometric forms. By contrast, DistilBERT can maintain a high degree of representational stability despite its shallower architecture, suggesting that compact transformer models can retain a significant amount of representational stability from deeper transformer models. ALBERT, on the other hand, always yields lower values of RSS, indicating that a lot of sharing of parameters affects the development of the geometric stability from layer to layer.

RoBERTa has a unique non-monotonic stability path. The RSSs are reduced by the intermediate layers, but after that, they are significantly boosted by the final representation layers, suggesting that intermediate layers specialize in features while the final layers represent higher-level semantics. This is a stark contrast to the progressive stability growth found in BERT, and underscores the effect of the pre-training tasks and architecture on the geometric organization of transformer representations. The layer-wise differences observed show that RSS is able to reveal structural properties of representation learning that are not visible from standard downstream performance measures.

It can be seen that the proposed metric is statistically stable across five different random seeds, as the confidence intervals are always narrow for all the transformer architectures tested. This suggests that the layer-wise RSS trajectories are statistically consistent and not due to chance sampling variability.

Yet another example of the usefulness of RSS is the cross-domain experiment that is also described. The transfer of BERT layer-12 representations from AG News to the *sci.med* domain reduces RSS from 5.455 to 2.042 with a resultant reduction in representation stability of 62.6%. This substantial drop shows that domain shift can severely affect the geometric quality of learned representations, despite keeping all the parameters of the pretrained model as is. Since it works directly with the intrinsic geometry of the embedding space, RSS is also unsupervised, meaning that it does not need to be trained on target domain labeled data or to be applied to downstream classifiers.

7 Conclusion

We showed that layer-wise geometric-spectral analysis of transformer text representations, using Representation Stability Score (RSS). By applying RSS to different layers of BERT, RoBERTa, DistilBERT, and ALBERT, it was shown that representation stability changes in architecture-specific manners when applied to SST-2 and IMDb benchmarks. The differences between BERT and RoBERTa stem from different optimization objectives and architecture during pretraining. BERT shows a mostly monotonic trajectory in its RSS as a function of depth, while RoBERTa shows a nonmonotonic trajectory. The statistical robustness and statistical repeatability of this proposed metric were further substantiated by the bootstrap confidence intervals across five independent random seeds. Moreover, the cross-domain experiment showed that RSS could also be used to capture representational degradation with domain shift, reducing by 62.6% in the cross-domain test of transferring BERT layer-12 representations from AG News to the *sci.med* domain. These results show that RSS could also be successfully generalized for characterizing representation geometry in the context of natural language processing with transformers. Going forward, the authors would like to explore the application of RSS to larger language models, instruction-tuned and fine-tuned transformers, multilingual

benchmarks, and even vision-language foundation models.

8 Limitations

In this paper, we test pretrained transformer models trained on frozen examples across sentiment classification datasets and a single cross-domain task. Future works will include fine-tuned large language models, multilingual transformers, and other downstream tasks to be validated and show the generalization of the proposed framework.

9 References

- [1] Mehmood, Esha, Safeera Batool, Ilyas Khan, and S. W. Teklu. 2026. "A Unified Geometric-Spectral Framework for Representation Stability Analysis in Self-Supervised Learning." Manuscript under review.
- [2] Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. "BERT Rediscovered the Classical NLP Pipeline." Proceedings of ACL, 4593-601.
- [3] Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. "What Does BERT Learn about the Structure of Language?" Proceedings of ACL, 3651-57.
- [4] Roy, Olivier, and Martin Vetterli. 2007. "The Effective Rank: A Measure of Effective Dimensionality." Proceedings of EUSIPCO, 606-10.
- [5] Trefethen, Lloyd N., and David Bau. 1997. Numerical Linear Algebra. SIAM.
- [6] Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. "A Theory of Learning from Different Domains." Machine Learning 79:151-175.
- [7] Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. "Attention Is All You Need." Advances in Neural Information Processing Systems (NeurIPS), 5998-6008.
- [8] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Proceedings of NAACL-HLT, 4171-4186.

- [9] Liu, Yinhan, Myle Ott, Naman Goyal, et al. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint arXiv:1907.11692.
- [10] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." NeurIPS Workshop on Energy Efficient Deep Learning.
- [11] Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations." International Conference on Learning Representations.
- [12] Socher, Richard, Alex Perelygin, Jean Wu, et al. 2013. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." Proceedings of EMNLP, 1631-1642.
- [13] Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." Proceedings of EMNLP Workshop BlackboxNLP, 353-355.
- [14] Maas, Andrew, Raymond Daly, Peter Pham, Dan Huang, Andrew Ng, and Christopher Potts. 2011. "Learning Word Vectors for Sentiment Analysis." Proceedings of ACL, 142-150.
- [15] Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. "Character-Level Convolutional Networks for Text Classification." Advances in Neural Information Processing Systems (NeurIPS), 649-657.
- [16] Lang, Ken. 1995. "Newsweeder: Learning to Filter Netnews." Proceedings of ICML, 331-339.
- [17] Wolf, Thomas, Lysandre Debut, Victor Sanh, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." Proceedings of EMNLP (System Demonstrations), 38-45.
- [18] Efron, Bradley, and Robert J. Tibshirani. 1994. An Introduction to the Bootstrap. Chapman & Hall.

