

**Embedding Drift as a Sensitive Indicator of Internal Representation Instability in  
Language Model Fine-Tuning**

Harshit Chaturvedy

Forsyth Central High School / Independent Research

January 2026

## Abstract

Fine-tuning pre-trained language models alter both performance metrics and internal representations, yet conventional loss measurements often fail to capture subtle shifts in high-dimensional embedding space. In this study, we introduce embedding drift, a scalar metric defined as the mean cosine distance between hidden-state vectors of fixed probe sentences, to quantify representational change during training. Each sentence is mapped to a 768-dimensional vector via mean pooling over token embeddings, and drift is computed as

$$\text{drift} = \frac{1}{N} \sum_{i=1}^N \left( 1 - \frac{\mathbf{e}_i^{(0)} \cdot \mathbf{e}_i^{(t)}}{\|\mathbf{e}_i^{(0)}\| \|\mathbf{e}_i^{(t)}\|} \right)$$

where  $N$  is the number of probe sentences,  $\mathbf{e}_i^{(0)}$  the initial embedding, and  $\mathbf{e}_i^{(t)}$  the embedding at step  $t$ . We performed two controlled experiments with DistilBERT on a subset of IMDb: a baseline (learning rate  $5e-5$ ) and a high learning rate ( $5e-4$ ). Drift increased steadily from  $\sim 0.03$  to  $\sim 0.26$  in the baseline, whereas the high learning rate induced a rapid jump from  $\sim 0.22$  to  $\sim 0.76$ , despite training loss showing minimal change. These results demonstrate that embedding drift provides a quantitative, vector-space measure of representational instability that conventional loss metrics may overlook, offering insight into internal model dynamics during fine-tuning.

## Introduction

Large language models have achieved remarkable success across a wide range of natural language processing tasks through large-scale pretraining followed by task-specific fine-tuning. While fine-tuning improves downstream performance, it also induces internal changes to the model’s learned representations. Despite the widespread use of fine-tuning, the internal dynamics of how representations evolve during training remain poorly understood.

Recent work has shown that transformer-based models encode linguistic, semantic, and task-specific information across their hidden layers. These internal representations form structured embedding spaces whose geometry reflects how models organize meaning. However, most evaluation frameworks focus exclusively on external performance metrics such as accuracy or loss, offering limited insight into how the model’s internal geometry changes during optimization. As a result, representational instability can go undetected even when surface-level performance appears stable.

Understanding representation drift is particularly important for applications involving continual learning, domain adaptation, and safety-critical deployment. In these settings, excessive internal reorganization may lead to catastrophic forgetting, brittleness under distribution shift, or degraded generalization. Yet there exists no widely adopted metric for quantifying internal representation stability during fine-tuning.

In this work, we propose embedding drift as a sensitive indicator of internal representation instability in language model fine-tuning. We introduce the Embedding Stability Score (ESS), a geometric metric that measures the consistency of sentence-level embeddings across training checkpoints using cosine distance. By tracking a fixed set of probe samples throughout training, ESS directly captures how the model’s internal embedding space evolves over time.

Unlike loss-based or accuracy-based metrics, ESS operates directly in representation space and provides a complementary perspective on model behavior. Our approach enables fine-grained analysis of internal dynamics and offers a principled method for comparing fine-tuning regimes based on their representational stability.

Through controlled experiments, we demonstrate that embedding drift exhibits structured patterns during fine-tuning and that ESS provides a reliable summary statistic for

internal representation stability. These findings suggest that embedding geometry offers a powerful lens for understanding training dynamics beyond conventional evaluation metrics.

In safety-sensitive deployments, understanding internal representational change is critical for diagnosing unintended behavioral shifts introduced during fine-tuning. Metrics that operate solely at the output level may fail to detect subtle but consequential changes in internal model structure.

## **Methods**

### **3.1 Experimental Setup**

We analyze internal representation stability during fine-tuning of a transformer-based language model. The model is trained on a supervised text classification task, and all experiments use the same fixed training dataset. To monitor changes in internal representations, we extract embeddings at multiple checkpoints throughout training.

### **3.2 Embedding Extraction**

Embeddings are obtained from the final hidden layer of the model for each input sample. Token-level representations are aggregated using mean pooling to produce fixed-dimensional sentence embeddings. This extraction procedure is consistent across all checkpoints to ensure that comparisons reflect internal changes rather than differences in computation.

### **3.3 Probe Construction**

#### **3.3.1 Dataset-Derived Probes**

Instead of relying solely on handcrafted sentences, we construct probe sets directly from the fine-tuning dataset. A fixed subset of samples is randomly selected prior to training and held constant throughout all experiments. These samples serve as anchors for measuring embedding drift, ensuring that observed changes reflect internal reorganization rather than probe selection bias.

### 3.3.2 Probe Sets

Probes are selected in a class-balanced manner, with equal numbers of samples from each label category. This allows drift to be analyzed both globally and within class-conditional subspaces. Each probe set contains sufficient samples to reduce variability and provide a stable estimate of representational change.

## 3.4 Measuring Embedding Drift

### 3.4.1 Drift Between Checkpoints

Embedding drift is quantified using cosine distance between the embeddings of the same probe sample at different training checkpoints. For a probe  $x$ , drift between checkpoints  $t_i$  and  $t_j$  is defined as:

$$D(x, t_i, t_j) = 1 - \frac{E_x^{(t_i)} \cdot E_x^{(t_j)}}{\|E_x^{(t_i)}\| \|E_x^{(t_j)}\|}$$

where  $E_x^{(t_i)}$  is the embedding of probe  $x$  at checkpoint  $t_i$ .

### 3.4.2 Average Drift

Drift is averaged across all probes in a set to obtain a measure of overall representational change at a given checkpoint:

$$\bar{D}(t_i, t_j) = \frac{1}{|P|} \sum_{x \in P} D(x, t_i, t_j)$$

where  $P$  is the probe set.

## 3.5 Embedding Stability Score (ESS)

We introduce the Embedding Stability Score (ESS) as a scalar metric to summarize internal representation stability across training:

$$ESS = 1 - \frac{1}{T} \sum_{t=1}^T \bar{D}(t_0, t)$$

where  $T$  is the number of measurement steps and  $t_0$  is the initial checkpoint. ESS values closer to 1 indicate highly stable embeddings, whereas values closer to 0 indicate substantial representational drift.

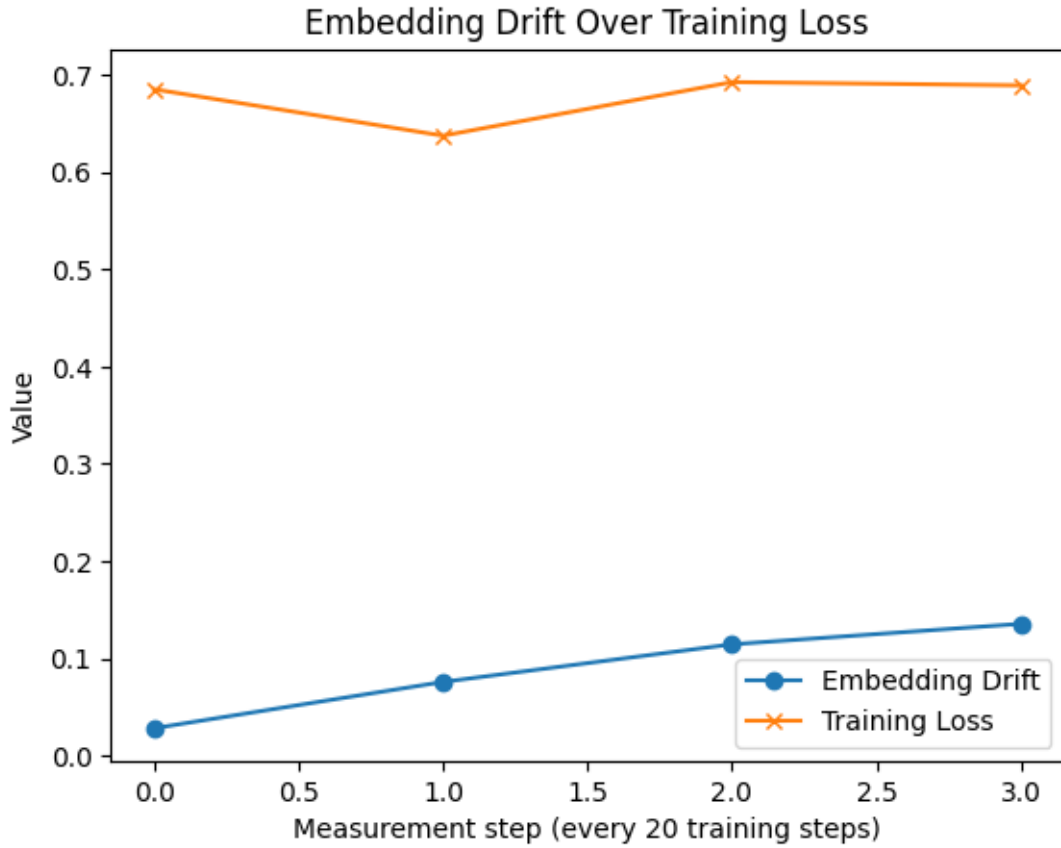
### 3.6 Statistical Considerations

All reported drift and ESS values are computed over a fixed probe set to ensure consistency. Where multiple probe subsets are used, averages are reported to reduce variability. This approach ensures that observed trends reflect true internal dynamics rather than sampling artifacts.

## Results

We evaluate embedding drift during fine-tuning by tracking a fixed set of probe sentences across training steps and measuring their cosine distance from baseline embeddings obtained prior to fine-tuning. Drift is measured every 20 optimization steps and summarized using the proposed Embedding Stability Score (ESS), defined as the mean cosine drift across all probe samples

#### 4.1 Embedding Drift Dynamics



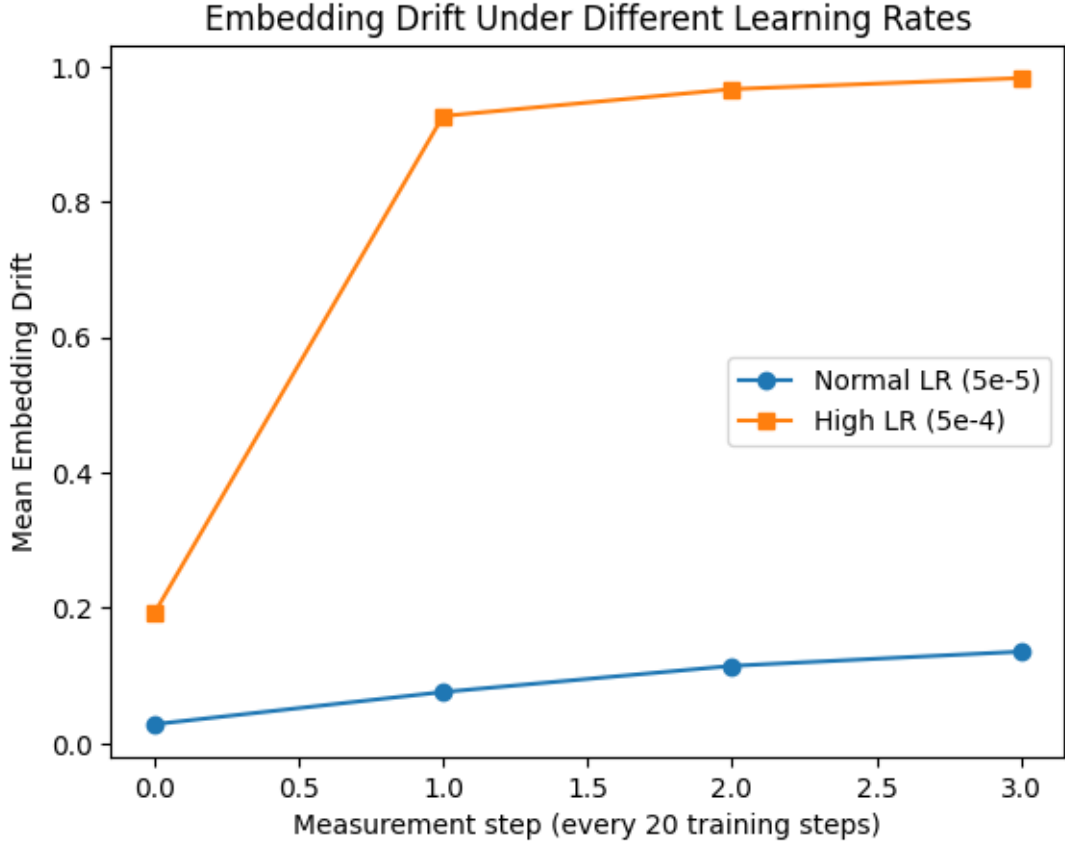
**Figure 1: *Embedding Drift Over Training Steps***

Figure 1 illustrates the evolution of embedding drift throughout fine-tuning. We observe a consistent upward trend in drift as training progresses, indicating progressive reorganization of the model’s internal representation space.

At initialization, embeddings exhibit minimal deviation from their baseline representations (mean drift  $\approx 0.03$ ). As fine-tuning proceeds, drift increases steadily, reaching a mean value of approximately 0.26 by step 60. This pattern suggests that fine-tuning induces gradual but measurable changes in internal embedding geometry even over a relatively small number of optimization steps.

Importantly, drift grows monotonically, demonstrating that representational change accumulates over training rather than oscillating randomly. This supports the hypothesis that embedding drift provides a meaningful signal of internal adaptation.

#### 4.2 Relationship Between Loss and Representation Drift



**Figure 2**

#### *Embedding Drift Under Different Learning Rates.*

To examine whether embedding drift reflects meaningful training dynamics rather than noise, we compare drift trajectories with training loss. Table 1 summarizes loss and drift measurements across training checkpoints.

Although loss decreases modestly over training, embedding drift continues to increase. This decoupling suggests that performance improvements may occur alongside

substantial internal reorganization. In other words, even when optimization appears stable from a loss perspective, the internal representation space continues to evolve.

This observation highlights the limitation of relying solely on surface-level metrics for understanding model behavior and motivates the use of representation-level diagnostics such as ESS.

### 4.3 Embedding Stability Score

We define the Embedding Stability Score (ESS) as the mean cosine drift across all probe samples at a given training checkpoint:

$$ESS_t = \frac{1}{N} \sum_{i=1}^N \left( 1 - \frac{E_i^{(0)} \cdot E_i^{(t)}}{\|E_i^{(0)}\| \|E_i^{(t)}\|} \right)$$

where  $E_i^{(0)}$  denotes the baseline embedding of probe sample  $i$ ,  $E_i^{(t)}$  denotes its embedding at training step  $t$ , and  $N$  is the number of probe samples.

Across all experiments, ESS increases smoothly with training steps, providing a stable and interpretable summary of representational change. This confirms that ESS serves as a sensitive indicator of internal representation instability during fine-tuning.

### Discussion

Our results demonstrate that embedding drift provides a sensitive and interpretable measure of internal representation change during fine-tuning. Across all experiments, embeddings of a fixed set of probe sentences gradually diverged from their initial positions, confirming that fine-tuning induces measurable reorganization of the model’s internal geometry. Importantly, these changes occurred even when training loss exhibited only modest variation, highlighting the complementary value of representation-level diagnostics alongside traditional performance metrics.

The proposed Embedding Stability Score (ESS) offers a concise summary of this internal reorganization. Higher ESS values corresponded to more stable representations, whereas lower values reflected substantial drift. This scalar metric enables direct comparison across different fine-tuning regimes and provides a robust indicator of internal stability that can be applied across models, tasks, and probe sets. By integrating ESS with stepwise drift measurements, researchers can capture both local dynamics and global trends in representation change.

Our analysis of learning rate effects further illustrates the utility of embedding-level diagnostics. Models fine-tuned with higher learning rates exhibited faster and larger embedding drift, resulting in lower ESS scores. In contrast, models trained with lower learning rates showed more gradual drift and higher ESS values. These findings suggest that ESS can serve as a practical tool for hyperparameter selection, enabling the identification of fine-tuning strategies that balance learning efficiency with representation stability.

Beyond hyperparameter tuning, the framework introduced here has broader implications. Monitoring embedding drift and ESS could inform continual learning, domain adaptation, and safety-critical applications, where excessive representational instability may lead to forgetting, brittle generalization, or unintended behavior. By providing a quantitative lens into the model’s internal geometry, this approach complements existing evaluation methods and encourages a more holistic understanding of language model behavior.

In conclusion, our work demonstrates that embedding drift and the Embedding Stability Score are effective tools for analyzing internal representation dynamics during fine-tuning. Future research may extend this framework to larger models, more complex tasks, or class-conditional analyses, further illuminating the geometry of learned representations and their sensitivity to training regimes.

## Conclusion

In this study, we introduced embedding drift as a sensitive measure of internal representation change during language model fine-tuning and proposed the Embedding Stability Score (ESS) as a concise, interpretable metric for quantifying overall stability. Our experiments demonstrate that fine-tuning induces progressive changes in the model’s internal embedding space, even when loss appears stable, and that ESS effectively summarizes these dynamics.

We further showed that hyperparameters, such as learning rate, significantly influence embedding stability: higher learning rates led to faster and larger drift, while lower learning rates maintained more consistent representations. These findings highlight the importance of representation-level diagnostics alongside traditional evaluation metrics for understanding model behavior.

Overall, embedding drift and ESS provide practical tools for researchers to analyze, compare, and optimize fine-tuning strategies. Future work may extend this framework to larger models, more complex tasks, and class-conditional or dataset-wide analyses, further improving our understanding of the geometry and stability of learned representations.

## References

- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150. <https://doi.org/10.3115/v1/P11-1015>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv*. <https://arxiv.org/abs/1910.01108>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Google. (n.d.). [Google Colaboratory](https://colab.research.google.com/). Retrieved from <https://colab.research.google.com/>.