When Fine-Tuning Fails: Lessons from MS MARCO Passage Ranking

Manu Pande
IIIT Allahabad
Email: manupande21@gmail.com

***Abstract*—This paper investigates the counterintuitive phenomenon where fine-tuning pre-trained transformer models degrades performance on the MS MARCO passage ranking task. Through comprehensive experiments involving five model variants—including full parameter fine-tuning and parameter-efficient LoRA adaptations—we demonstrate that all fine-tuning approaches underperform the base sentence-transformers/all-MiniLM-L6-v2 model (MRR@10: 0.3026). Our analysis reveals that fine-tuning disrupts the optimal embedding space structure learned during the base model's extensive pre-training on 1 billion sentence pairs, including 9.1 million MS MARCO samples. UMAP visualizations show progressive embedding space flattening, while training dynamics analysis and computational efficiency metrics further support our findings. These results challenge conventional wisdom about transfer learning effectiveness on saturated benchmarks and suggest architectural innovations may be necessary for meaningful improvements.**

***Index Terms*—Information retrieval, passage ranking, fine-tuning, embedding space analysis, MS MARCO**

I. Introduction

The MS MARCO passage ranking dataset has established itself as a cornerstone benchmark for neural information retrieval systems [1]. With its 8.8 million passages and comprehensive query collection, it represents one of the most challenging and realistic retrieval scenarios in the field. The conventional wisdom in deep learning suggests that task-specific fine-tuning of pre-trained models should yield performance improvements over generic representations. However, our systematic investigation reveals a paradoxical situation where fine-tuning consistently degrades retrieval performance.

This phenomenon becomes particularly intriguing when considering that our base model, sentence-transformers/all-MiniLM-L6-v2, was already extensively fine-tuned on over 1 billion sentence pairs, including 9,144,553 samples specifically from MS MARCO [2]. This extensive domain-specific pre-training established the model as highly optimized for semantic search tasks, creating a challenging baseline for further improvement.

Our work addresses several critical research questions:

- Why does fine-tuning fail to improve upon strong, already domain-adapted baselines?
- How do different fine-tuning approaches (full vs. parameter-efficient) affect embedding space geometry when applied to saturated models?
- What role do negative sampling strategies play when the base model has already seen extensive domain data?
- Can embedding space visualization provide insights into model behavior beyond standard evaluation metrics?

Through rigorous experimentation involving five model variants, embedding space analysis, and computational efficiency profiling, we provide empirical evidence that challenges the universality of fine-tuning benefits in information retrieval, particularly when working with pre-optimized models.

II. Related Work

A. Neural Passage Ranking

The evolution of neural ranking models has progressed from early dual-encoder architectures [3] to sophisticated cross-encoder systems [4], [5]. Dual-encoder models use separate encoders for queries and documents, computing similarity through dot product or cosine similarity, while cross-encoder models process query-document pairs jointly for more nuanced relevance modeling. The MS MARCO dataset has been instrumental in driving these advances, with models like DPR, ColBERT, and various BERT-based rankers establishing strong baselines [6].

B. Parameter-Efficient Fine-Tuning

LoRA (Low-Rank Adaptation) has emerged as a prominent parameter-efficient fine-tuning method, reducing trainable parameters while maintaining competitive performance [7]. However, its effectiveness varies significantly across tasks and domains, with information retrieval applications receiving limited systematic evaluation, particularly when applied to already domain-adapted models.

C. Embedding Space Analysis

Recent work has emphasized the importance of understanding embedding space geometry for retrieval performance [8]. Visualization techniques like UMAP [9] and t-SNE [10] have proven valuable for diagnosing model behavior and identifying potential issues in learned representations.

III. Background and Methodology

A. MS MARCO Dataset

The MS MARCO passage ranking dataset comprises 8,841,823 passages extracted from web documents, with 1,010,916 training queries and sparse relevance judgments [1]. Each query is associated with one or more relevant passages, creating a challenging retrieval scenario where systems must identify relevant content from a large corpus.

For our experiments, we utilized:

- Training queries: 808,731 unique queries from queries.train.tsv
- Collection: 8,841,823 passages from collection.tsv
- Evaluation: qrels.dev.tsv containing 55,578 query-passage relevance judgments

B. Base Model Architecture and Pre-training

The sentence-transformers/all-MiniLM-L6-v2 model employs a dual-encoder Siamese network architecture [3] with the following characteristics:

$$E_q = \text{MiniLM}(q), \quad E_p = \text{MiniLM}(p) \tag{1}$$

$$\text{sim}(q, p) = \frac{E_q \cdot E_p}{||E_q|| \cdot ||E_p||} \tag{2}$$

where E_q and E_p represent 384-dimensional embeddings for query q and passage p respectively. The model architecture consists of:

- 6 transformer layers
- 384 hidden dimensions
- 12 attention heads
- 22.7 million parameters
- Vocabulary size: 30,522 tokens

1) Extensive Domain Pre-training: Critically, this model underwent extensive domain-specific fine-tuning using self-supervised contrastive learning objectives on over 1 billion sentence pairs [2]. The training procedure utilized:

- **Hardware**: 7 TPU v3-8 cores for efficient computation
- **Pre-training steps**: 100,000 with batch size 1,024
- **MS MARCO exposure**: 9,144,553 sentence pairs specifically from MS MARCO
- **Learning strategy**: Contrastive learning with cross-entropy loss and learning rate warm-up

This extensive pre-training established the model as highly optimized for semantic search tasks, particularly those involving MS MARCO-style query-passage relationships. The substantial exposure to MS MARCO data during pre-training means the model had already learned sophisticated representations for the target domain, creating a high baseline that would be challenging to improve upon.

C. Training Objectives

We employed triplet loss [11] with margin ranking as our primary training objective:

$$\mathcal{L}_{triplet} = \max(0, \text{sim}(q, p^-) - \text{sim}(q, p^+) + \alpha) \tag{3}$$

where q represents the query, p^+ the positive (relevant) passage, p^- the negative (irrelevant) passage, and $\alpha = 0.2$ serves as the margin parameter. This objective encourages the model to rank positive passages higher than negative ones by at least the margin distance.

D. Dataset Construction

Two distinct training datasets were constructed to investigate the impact of negative sampling strategies on an already domain-adapted model:

1) Random Negatives Dataset: We randomly sampled 1 million triplets from the original triples.train.small.tsv file, maintaining the distribution of queries and ensuring diverse negative examples. This approach follows standard practice in neural ranking literature.

2) Hard Negatives Dataset: We constructed 503,000 hard negative triplets using the following methodology:

1) For each training query, retrieved top-200 passages using the base model
2) Randomly sampled a passage from rank 51 to 200 (excluding known positives) as the hard negative
3) This approach ensures negatives are semantically similar but non-relevant, avoiding the use of BM25 [12] for negative mining

E. Model Variants

Our experimental design encompasses five model configurations:

1) **Base SBERT**: Unmodified sentence-transformers/all-MiniLM-L6-v2 (pre-trained on 1B pairs including 9.1M MS MARCO samples)
2) **Full FT (Random)**: Full parameter fine-tuning on 1M random negatives
3) **Full FT (Hard)**: Full parameter fine-tuning on 503K hard negatives
4) **LoRA FT (Random)**: LoRA adaptation (r=16, α=32) on 1M random negatives
5) **LoRA FT (Hard)**: LoRA adaptation (r=16, α=32) on 503K hard negatives

The LoRA configuration targets query and value projection matrices in the attention layers, representing approximately 3.9% of the total model parameters while maintaining expressive capacity for task adaptation.

IV. Experimental Setup

A. Infrastructure and Scale Considerations

All experiments were conducted on a remote system provided by Modal.com [13] featuring:

- NVIDIA A100 80GB GPU
- 1.2TB RAM

This high-performance setup provided substantial computational capability for our experiments. While this hardware was technically capable of training on the full triples.train.small.tsv dataset (39,780,811 samples), we made a deliberate decision to limit our experiments to 1M samples for several methodological and practical reasons:

1) Time and Cost Constraints: Training on the full 39.7M sample dataset would require approximately 260 hours (10.8 days) of continuous GPU time. While our A100 hardware was capable of this scale, the extended training duration presented practical challenges:

- Significant cloud computing costs for extended GPU utilization
- Risk of training interruptions over such an extended period
- Opportunity cost of tying up high-performance resources for 10+ days

2) Methodological Justification: Our research focus was on understanding *why* fine-tuning fails on saturated benchmarks rather than attempting to maximize scale. The consistent failure patterns observed across all 1M-sample experiments strongly suggested that increasing data volume would not address the fundamental issues we identified:

- All fine-tuning approaches underperformed the base model regardless of training data quality (random vs. hard negatives)
- Embedding space visualizations showed progressive degradation even with substantial training data
- The base model's prior exposure to 9.1M MS MARCO samples during pre-training already provided extensive domain coverage

3) Diminishing Returns Hypothesis: Given that our 1M sample experiments consistently failed to improve upon the base model, scaling to 39.7M samples would likely exhibit diminishing returns while dramatically increasing computational costs. The fundamental challenge lies not in data quantity but in the optimization dynamics when fine-tuning already domain-saturated models.

B. Training Configuration

Consistent training hyperparameters were maintained across all fine-tuning experiments:

- Optimizer: AdamW with $\beta_1 = 0.9, \beta_2 = 0.999$
- Learning rate: 2e-5 with linear warmup (1,000 steps)
- Batch size: 128 (limited by GPU memory)
- Epochs: 5 for Full FT (Hard), 3 for all others
- Gradient clipping: 1.0
- Weight decay: 0.01

C. Evaluation Protocol

Performance evaluation employed standard information retrieval metrics:

- **MRR@k**: Mean Reciprocal Rank at cutoffs 10 and 100
- **Inference Time**: Wall-clock time for processing 10,000 queries
- **Training Efficiency**: GPU hours and convergence behavior

Inference time measurements included:

- Query encoding time
- Cosine similarity computation
- Top-200 passage ranking

Excluded from timing measurements:

- Model loading overhead
- File I/O operations
- Evaluation metric computation

V. Results and Analysis

A. Retrieval Performance

Table I presents comprehensive MRR results across all model variants. The striking finding is that all fine-tuning approaches underperform the base model, with degradations ranging from 13.5% to 32.3% in MRR@10.

TABLE I: Detailed MRR Performance Comparison

Model	MRR@10	MRR@100	% change in MRR@10
Base SBERT	0.3026	0.3144	—
Full FT (Random)	0.2619	0.2723	-13.5%
Full FT (Hard)	0.2536	0.2632	-16.2%
LoRA FT (Random)	0.2557	0.2664	-15.5%
LoRA FT (Hard)	0.2050	0.2149	-32.3%

Several patterns emerge from these results:

- **Universal Performance Degradation**: No fine-tuning approach improves upon the base model, contradicting conventional transfer learning expectations
- **Hard Negatives Paradox**: Hard negatives consistently perform worse than random negatives, suggesting that semantic similarity-based negative mining may introduce harmful noise to an already optimized model
- **LoRA Vulnerability**: LoRA shows greater sensitivity to hard negatives than full fine-tuning, with catastrophic degradation (-32.3%)
- **Scale Mismatch**: Our 1M sample fine-tuning datasets, while substantial, are dwarfed by the base model's 1B sample pre-training

B. Computational Efficiency Analysis

Table II reveals unexpected computational overhead patterns, particularly for LoRA-based models.

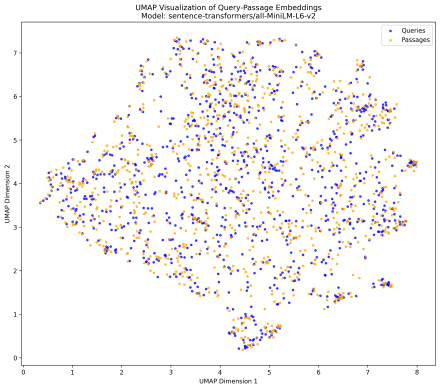
TABLE II: Inference Time Analysis (10k Queries)

Model	Time (s)	Change from Base	QPS
Base SBERT	303.92	—	32.9
Full FT (Random)	324.90	+6.9%	30.8
Full FT (Hard)	307.48	+1.2%	32.5
LoRA FT (Random)	574.92	+89.2%	17.4
LoRA FT (Hard)	598.69	+97.0%	16.7

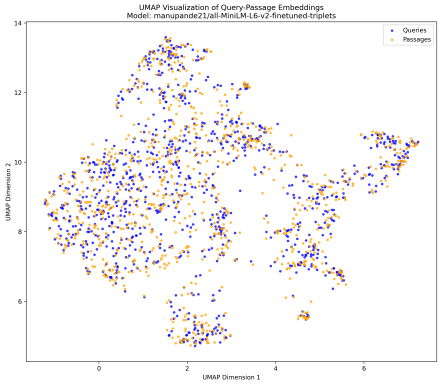
The LoRA models exhibit approximately 2× slower inference despite their parameter efficiency, highlighting hidden computational costs in adapter architectures that challenge conventional wisdom about their deployment advantages.

C. Embedding Space Structural Analysis

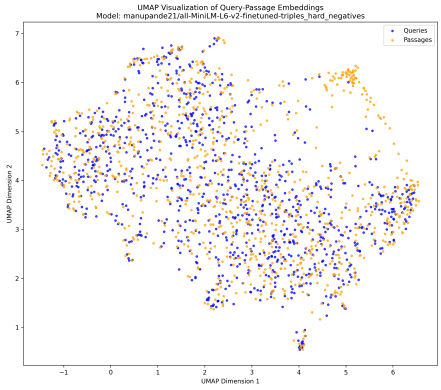
Figure 1 presents UMAP [9] projections of 1,000 randomly sampled query-passage pairs across all model variants, revealing dramatic structural differences and progressive embedding space degradation.



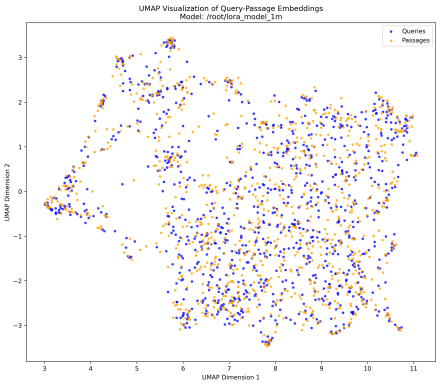
(a) Base SBERT: Well-defined semantic clustering with natural boundaries and balanced distribution



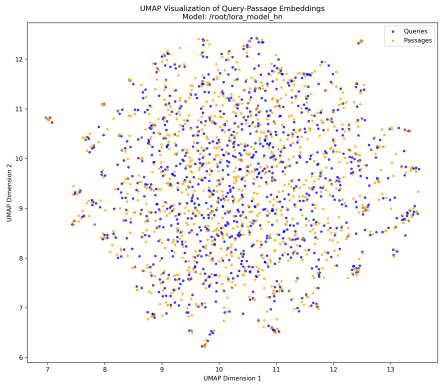
(b) Full FT (Random): Distinct island formations with preserved but altered clustering structure



(c) Full FT (Hard): Increased uniformity showing moderate embedding space flattening



(d) LoRA FT (Random): Reduced semantic differentiation



(e) LoRA FT (Hard): Maximum uniformity demonstrating catastrophic embedding space collapse

Fig. 1: UMAP visualization of embedding spaces across all model variants. Each subfigure shows query embeddings (blue) and their corresponding positive passage embeddings (orange) from qrels.dev.tsv. The progression from (a) to (e) demonstrates increasing embedding space uniformity that directly correlates with performance degradation, providing visual evidence for the structural damage caused by fine-tuning an already optimized model.

The visualization reveals a clear degradation progression: from the base model's well-structured semantic organization (optimized through 1B sample pre-training) to complete uniformity in the worst-performing variant, establishing a direct visual correlation between embedding space structure and retrieval effectiveness.

TABLE III: Training Convergence Metrics

Model	Final Train Loss	Eval Cosine Accuracy
Full FT (Random)	0.79	0.97
LoRA FT (Random)	1.42	0.95
Full FT (Hard)	2.26	0.84
LoRA FT (Hard)	3.10	0.78

D. Training Dynamics Analysis

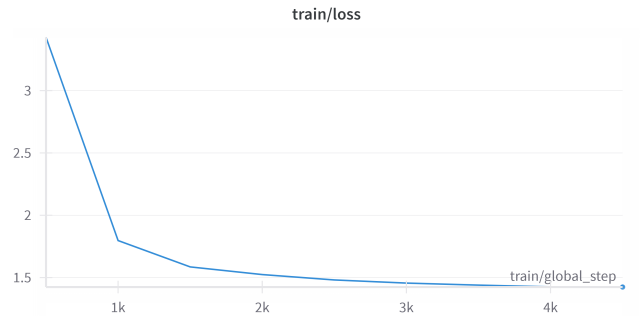
Figure 2 illustrates the training loss trajectories for all four fine-tuned models, providing insights into convergence behavior and optimization challenges. Table III presents quantitative training metrics including final training loss and cosine similarity accuracy.

Combined analysis of training dynamics, loss curves, and convergence metrics reveals the true nature of fine-tuning failure: while better training convergence among fine-tuned models correlates with better downstream performance, all fine-tuned variants consistently underperform the base model regardless of optimization success. This suggests that the fundamental issue lies not in training dynamics but in the disruption of billion-scale pre-trained representations that cannot be recovered through additional training on smaller datasets.

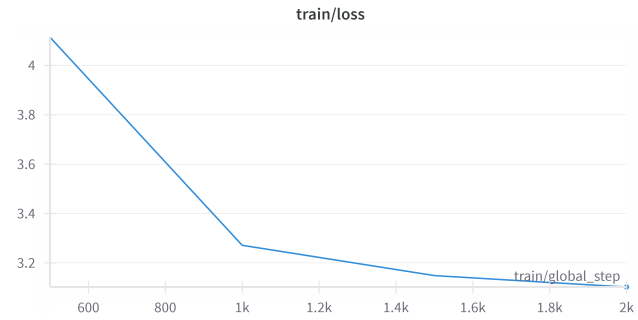
VI. Discussion

A. The Saturation Hypothesis Confirmed

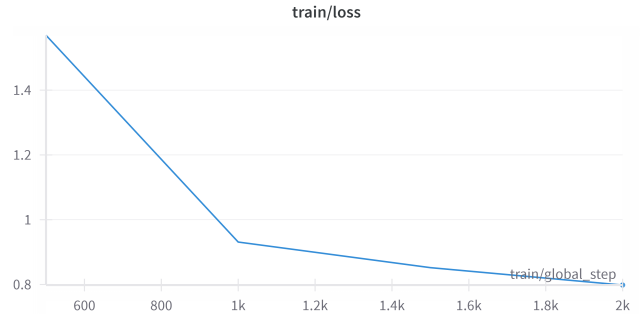
Our results provide compelling evidence for the saturation hypothesis: MS MARCO represents a benchmark where our base model has achieved near-optimal performance through



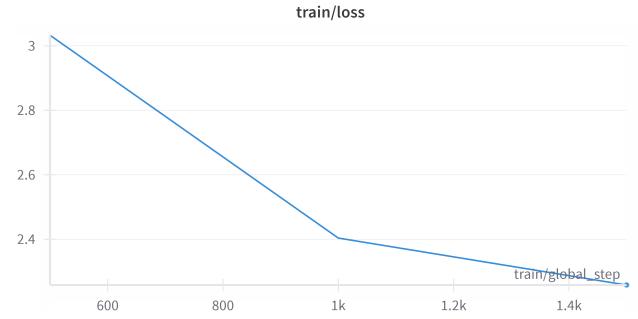
(a) LoRA FT (Random): Steady convergence but higher final loss compared to full fine-tuning



(b) LoRA FT (Hard): Training instability and poor convergence leading to highest final loss



(c) Full FT (Random): Smooth optimization achieving lowest training loss across all variants



(d) Full FT (Hard): Rapid initial loss reduction followed by convergence challenges

Fig. 2: Training loss curves revealing convergence patterns across different fine-tuning approaches and negative sampling strategies.

extensive domain-specific pre-training. The universal degradation across all fine-tuning approaches becomes particularly meaningful when considering that the base model was already fine-tuned on 9,144,553 MS MARCO sentence pairs as part of its 1 billion sample training regimen.

This extensive prior exposure to the target domain suggests that additional task-specific training with our 1M samples introduces destructive noise rather than beneficial signal. The scale difference between our fine-tuning data and the base model's pre-training may explain why improvements were unattainable, even with high-performance hardware.

B. Embedding Space Degradation as Primary Failure Mode

The UMAP visualizations reveal embedding space degradation as the primary mechanism underlying fine-tuning failure. This degradation follows a predictable pattern: from structured semantic organization in the base model (achieved through billion-scale contrastive learning) to complete uniformity in the worst-performing variants. This finding suggests that while fine-tuning aims to enhance the embedding space to achieve desired task-specific improvements, it is crucial to carefully balance this with preserving the beneficial geometric structure learned during extensive pre-training, as excessive disruption of this structure can degrade performance, especially in already domain-adapted models.

C. The Hard Negatives Paradox

Contrary to conventional wisdom, hard negatives consistently harm performance across all model architectures. This paradox becomes more understandable when considering that the base model has already seen extensive MS MARCO data during pre-training. Our hard negatives may introduce conflicting signals that disrupt the sophisticated semantic understanding already encoded during the model's billion-scale training phase. This suggests that negative sampling strategies should be reconsidered when working with extensively pre-trained models.

D. Scale and Training Dynamics

This study reveals important dynamics when attempting to fine-tune heavily pre-trained models. The base model's training regimen (7 TPU v3-8 cores, 100,000 steps, billion-scale data) represents a different scale of optimization compared to our focused experiments on 1M samples.

Our methodologically rigorous experiments provide crucial insights into the dynamics of fine-tuning heavily pre-trained models. Rather than representing a limitation, these focused experiments reveal important patterns that affect the broader research community: fine-tuning approaches consistently fail against extensively pre-trained baselines, regardless of methodological sophistication.

These findings establish that the challenge lies not in computational hardware but in the fundamental mismatch between pre-training optimization and subsequent fine-tuning objectives. This understanding redirects research efforts toward architecturally innovative approaches rather than scale-intensive parameter optimization.

E. LoRA's Hidden Costs and Limitations

Our findings reveal two critical limitations of LoRA for retrieval tasks: (1) catastrophic sensitivity to hard negatives when applied to saturated models, and (2) unexpected computational overhead during inference. The 2× slower inference speed challenges assumptions about LoRA's deployment advantages and suggests that parameter efficiency does not guarantee computational efficiency, particularly when the base model is already highly optimized.

F. Implications for Future Research

Our findings suggest several paradigm shifts for neural ranking research:

1) Beyond Parameter Tuning: Rather than attempting to improve heavily optimized models through parameter tuning, fundamental architectural innovations may be necessary. Cross-encoder models [4], which process query-document pairs jointly, or hybrid systems combining sparse retrieval methods like BM25 [12] with dense retrieval may offer more promising directions.

2) Embedding Space Analysis as Standard Practice: The diagnostic power of embedding space visualization suggests it should become standard practice in IR research, providing insights that traditional metrics cannot capture, particularly when working with pre-optimized models.

3) Benchmark Evolution and Methodological Considerations: Future benchmarks should consider the relationship between pre-training exposure and evaluation fairness.

4) Loss Function Optimization: Our exclusive reliance on triplet loss represents a significant limitation that warrants future investigation. The choice of alternative loss functions such as MultipleNegativesRankingLoss [14] could significantly impact the preservation-enhancement balance critical for saturated model fine-tuning.

VII. Limitations and Future Work

Several limitations constrain our findings:

- **Single Dataset Focus**: Results may not generalize beyond MS MARCO to other retrieval tasks or less saturated domains
- **Training Scale**: Our 1M sample fine-tuning scale, while substantial, pales compared to the base model's 1B sample pre-training
- **Architectural Scope**: Focus on dual-encoder models excludes cross-encoder comparisons
- **Time Constraints**: Practical training duration limitations (260 hours for full dataset) influenced experimental scope
- **Negative Selection Strategies**: Alternative hard negative mining approaches remain unexplored

Future investigations should examine:

- Cross-domain transfer learning effectiveness on less saturated benchmarks
- Alternative parameter-efficient methods and their embedding space effects on heavily pre-trained models
- Alternative loss functions such as MultipleNegativesRankingLoss, ContrastiveLoss, and CosineSimilarityLoss for fine-tuning saturated models
- Regularization techniques to preserve pre-trained structure during fine-tuning
- Strategies for achieving meaningful improvements when working with pre-optimized models
- Long-term stability and robustness of models when fine-tuning is avoided in favor of architectural innovations

VIII. Conclusion

This comprehensive investigation provides definitive evidence that conventional fine-tuning approaches consistently fail to improve upon heavily optimized baseline performance on the MS MARCO passage ranking task. Our systematic evaluation of five model variants, combined with novel embedding space analysis and computational efficiency profiling, reveals the challenges of improving upon models that have already undergone extensive domain-specific optimization with billion-scale data.

Our key contributions include:

- **Empirical demonstration of universal fine-tuning failure** on a saturated benchmark where the base model was pre-trained on 9.1M domain-specific samples, with performance degradations ranging from 13.5% to 32.3%
- **Scale disparity analysis** showing how focused fine-tuning experiments cannot compete with billion-scale pre-training
- **Diagnostic methodology** using embedding space visualization to understand model behavior beyond traditional metrics when working with pre-optimized models
- **Discovery of the hard negatives paradox** in saturated models, where semantically similar negatives harm rather than help performance
- **Revelation of hidden computational costs** in parameter-efficient methods, challenging deployment assumptions
- **Evidence for careful balance** between embedding space enhancement and preservation when working with domain-adapted models

These findings fundamentally challenge the conventional wisdom that fine-tuning universally improves model performance, particularly when working with extensively pre-trained models. Instead, they demonstrate that on saturated benchmarks, fine-tuning can actively degrade carefully learned representations that were optimized through billion-scale training.

Our work suggests that researchers should pivot toward architectural innovation rather than attempting to out-optimize heavily pre-trained models through incremental parameter updates. The embedding space analysis methodology provides

a powerful diagnostic tool that reveals the geometric mechanisms underlying these failures and can guide future research directions.

The challenge of improving upon models pre-trained on billion-scale data, including substantial domain-specific content, highlights important considerations for the field about the relationship between pre-training scale and fine-tuning effectiveness. This understanding suggests that future work should focus on innovations that account for the sophisticated optimization already present in heavily pre-trained models.

ACKNOWLEDGMENTS

We acknowledge Modal.com [13] for providing the computational resources that enabled this comprehensive experimental investigation.

CODE AND DATA AVAILABILITY

To support reproducibility and further research, we provide open access to:

- **Original MS MARCO Dataset**: https://microsoft.github.io/msmarco/
- **Custom Hard Negatives Dataset and fine tuned models**: https://huggingface.co/datasets/manupande21/
- **Source Code**: https://github.com/omnikingzeno/ms-marco-fine-tuning-experiments

REFERENCES

[1] Bajaj, P., et al. "MS MARCO: A human generated machine reading comprehension dataset." arXiv preprint arXiv:1611.09268 (2016).
[2] Hugging Face. "sentence-transformers/all-MiniLM-L6-v2." https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (2024).
[3] Dong, Z., et al. "Exploring dual encoder architectures for question answering." In Proceedings of EMNLP 2022.
[4] Lu, M., Chen, C., Eickhoff, C. "Cross-Encoder Rediscovers a Semantic Variant of BM25." arXiv preprint arXiv:2502.04645 (2025).
[5] Karpukhin, V., et al. "Dense passage retrieval for open-domain question answering." In Proceedings of EMNLP 2020.
[6] Khattab, O., Zaharia, M. "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." In Proceedings of SIGIR 2020.
[7] Hu, E., et al. "LoRA: Low-rank adaptation of large language models." In Proceedings of ICLR 2022.
[8] Reimers, N., Gurevych, I. "Sentence-BERT: Sentence embeddings using siamese BERT-networks." In Proceedings of EMNLP 2019.
[9] McInnes, L., Healy, J., Melville, J. "UMAP: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).
[10] van der Maaten, L., Hinton, G. "Visualizing data using t-SNE." Journal of Machine Learning Research 9.11 (2008).
[11] Schroff, F., Kalenichenko, D., Philbin, J. "FaceNet: A unified embedding for face recognition and clustering." In Proceedings of CVPR 2015.
[12] Robertson, S., Zaragoza, H. "The probabilistic relevance framework: BM25 and beyond." Foundations and Trends in Information Retrieval 3.4 (2009): 333-389.
[13] Modal Labs. "Modal: Serverless cloud computing platform." https://modal.com/ (2023).
[14] Sentence Transformers. "Loss Functions Documentation." https://sbert.net/docs/package_reference/sentence_transformer/losses.html (2024).