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

A thesis submitted in partial fulfillment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

By:-
MANU PANDE

Enrollment No.
MML2023005

Under the Supervision of
Dr. Muneendra Ojha

to the
Department of Information Technology

भारतीय सूचना प्रौद्योगिकी संस्थान, इलाहाबाद
Indian Institute of Information Technology, Allahabad

June, 2025

भारतीय सूचना प्रौद्योगिकी संस्थान इलाहाबाद
Indian Institute of Information Technology Allahabad
An Institute of National Importance by Act of Parliament
Deoghat Jhalwa, Prayagraj 211015 (U.P.) India
Ph: 0532-2922025, 2922067; Web: www.iiita.ac.in; Email: contact@iiita.ac.in

CANDIDATE DECLARATION

I hereby declare that work presented in the report entitled "**When Fine-Tuning Fails: Lessons from MS MARCO Passage Ranking**", submitted towards the fulfillment of MASTER'S THESIS report of M.Tech at Indian Institute of Information Technology Allahabad, is an authenticated original work carried out under supervision of **Dr. Muneendra Ojha**. Due Acknowledgements have been made in the text to all other material used. the project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Manu Pande - MML2023005

भारतीय सूचना प्रौद्योगिकी संस्थान इलाहाबाद
Indian Institute of Information Technology Allahabad
An Institute of National Importance by Act of Parliament
Deoghat Jhalwa, Prayagraj 211015 (U.P.) India
Ph: 0532-2922025, 2922067; Web: www.iiita.ac.in; Email: contact@iiita.ac.in

CERTIFICATE FROM SUPERVISORS

It is certified that the work contained in the thesis titled "**When Fine-Tuning Fails: Lessons from MS MARCO Passage Ranking**" by **Manu Pande** has been carried out under supervision of **Dr. Muneendra Ojha** and that this work has not been submitted elsewhere for a degree.

Dr. Muneendra Ojha
Department of Information Technology
IIIT Allahabad

भारतीय सूचना प्रौद्योगिकी संस्थान इलाहाबाद
Indian Institute of Information Technology Allahabad
An Institute of National Importance by Act of Parliament
Deoghat Jhalwa, Prayagraj 211015 (U.P.) India
Ph: 0532-2922025, 2922067; Web: www.iiita.ac.in; Email: contact@iiita.ac.in

CERTIFICATE OF APPROVAL

This thesis entitled **When Fine-Tuning Fails: Lessons from MS MARCO Passage Ranking** by **Manu Pande** (MML2023005) is approved for the degree of Master's thesis at IIIT Allahabad It is understood that by this approval, the undersigned does not necessarily endorse or approve any statement made, opinion expressed, or conclusion drawn therein but approves the thesis only for the purpose for which it is submitted."

Signature and name of the committee members (on final examination and approval of the thesis):

1. Dr. Muneendra Ojha
2. Dr. [Committee Member 2]
3. Dr. [Committee Member 3]

Dean(A&R)

भारतीय सूचना प्रौद्योगिकी संस्थान इलाहाबाद
Indian Institute of Information Technology Allahabad
An Institute of National Importance by Act of Parliament
Deoghat Jhalwa, Prayagraj 211015 (U.P.) India
Ph: 0532-2922025, 2922067; Web: www.iiita.ac.in; Email: contact@iiita.ac.in

ACKNOWLEDGEMENT

I am thankful to my project supervisor, **Dr. Muneendra Ojha** for the guidance, support, and invaluable feedback throughout the research process. Their expertise, encouragement, and patience have been instrumental in the completion of this thesis.

I am grateful to the faculty and staff of the Department of Information Technology at IIIT Allahabad for their support throughout my research journey. I would also like to thank the Indian Institute of Information Technology, Allahabad for providing the necessary resources and facilities to conduct this research.

I acknowledge Modal.com for providing the computational resources that enabled this comprehensive experimental investigation. Finally, I am deeply grateful to my family and friends for their unwavering support and encouragement throughout this academic pursuit.

Manu Pande - MML2023005

ABSTRACT

This thesis 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.

The key contributions include empirical demonstration of universal fine-tuning failure on saturated benchmarks, scale disparity analysis showing how focused fine-tuning experiments cannot compete with billion-scale pre-training, and diagnostic methodology using embedding space visualization to understand model behavior beyond traditional metrics when working with pre-optimized models. This work provides crucial insights for the information retrieval community about the limitations of conventional fine-tuning approaches on heavily optimized baseline models.

Table Of Contents

List of Figures

List of Tables

List of Abbreviations

MS MARCO	Microsoft MAchine Reading COmprehension
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
LoRA	Low-Rank Adaptation
BERT	Bidirectional Encoder Representations from Transformers
SBERT	Sentence-BERT
ColBERT	Contextualized Late Interaction over BERT
FT	Fine-Tuning
UMAP	Uniform Manifold Approximation and Projection
t-SNE	t-distributed Stochastic Neighbor Embedding
IR	Information Retrieval
NLP	Natural Language Processing
TPU	Tensor Processing Unit
GPU	Graphics Processing Unit
QPS	Queries Per Second
BM25	Best Matching 25
DPR	Dense Passage Retrieval

Chapter 1

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 was already extensively pre-trained on over 1 billion sentence pairs, including substantial MS MARCO data [7], establishing an already-optimized baseline for semantic search tasks.

1.1 Research Questions

Our work addresses several critical research questions:

- Why does fine-tuning fail to improve upon the 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?

1.2 Contributions

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.

Our key contributions include:

- **Empirical demonstration of universal fine-tuning failure** on a saturated benchmark where the base model was extensively pre-trained on domain-specific samples, with significant performance degradations across all variants
- **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

1.3 Thesis Organization

This thesis is organized as follows:

- **Chapter 2** reviews related work in neural passage ranking, parameter-efficient fine-tuning, and embedding space analysis
- **Chapter 3** identifies the research gap and presents our proposed investigation approach
- **Chapter 4** details our experimental methodology, dataset construction, and model configurations
- **Chapter 5** presents comprehensive results including performance metrics, computational efficiency analysis, and embedding space visualizations
- **Chapter 6** summarizes findings and discusses future research directions

Chapter 2

Literature Review

This chapter provides a comprehensive review of the literature relevant to our investigation of fine-tuning failures in neural passage ranking. We organize our review around three key areas: neural passage ranking architectures, parameter-efficient fine-tuning methods, and embedding space analysis techniques.

2.1 Neural Passage Ranking

2.1.1 Evolution of Ranking Architectures

The evolution of neural ranking models has progressed from early dual-encoder architectures [4] to sophisticated cross-encoder systems [11, 8]. This progression represents a fundamental shift in how neural systems approach the information retrieval task.

Dual-Encoder Models: These architectures use separate encoders for queries and documents, computing similarity through dot product or cosine similarity. The key advantage of dual-encoder models lies in their computational efficiency during inference, as document embeddings can be pre-computed and indexed. However, they sacrifice some accuracy due to the lack of query-document interaction during encoding.

Cross-Encoder Models: These systems process query-document pairs jointly, enabling more nuanced relevance modeling through attention mechanisms [18] that allow fine-grained interaction between query and document tokens. While computationally expensive, cross-encoders typically achieve superior ranking performance.

2.1.2 MS MARCO and Benchmark Evolution

The MS MARCO dataset has been instrumental in driving advances in neural ranking [1]. Models like DPR (Dense Passage Retrieval), ColBERT, and various BERT-based rankers [3] have established strong baselines on this benchmark [9]. The dataset's scale and realistic query distribution have made it a standard evaluation framework for the information retrieval community.

However, the extensive use of MS MARCO in model pre-training has created a unique challenge: many contemporary models have already been exposed to substantial portions of the evaluation data during their pre-training phase, potentially leading to benchmark saturation.

2.2 Parameter-Efficient Fine-Tuning

2.2.1 LoRA: Low-Rank Adaptation

LoRA (Low-Rank Adaptation) has emerged as a prominent parameter-efficient fine-tuning method [6]. The core insight behind LoRA is that the weight updates during fine-tuning often have a low "intrinsic rank," meaning they can be approximated by low-rank matrices.

Mathematically, LoRA represents weight updates as:

$$W' = W + \Delta W = W + BA \tag{2.1}$$

where $W \in \mathbb{R}^{d \times k}$ is the original weight matrix, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are trainable low-rank matrices, and $r \ll \min(d, k)$ is the rank.

2.2.2 Effectiveness Across Domains

While LoRA has shown success in natural language processing tasks, its effectiveness varies significantly across tasks and domains. Information retrieval applications have received limited systematic evaluation, particularly when applied to already domain-adapted models. This gap in the literature motivates our investigation into LoRA's behavior on saturated benchmarks.

2.3 Embedding Space Analysis

2.3.1 Geometric Understanding of Embeddings

Recent work has emphasized the importance of understanding embedding space geometry for retrieval performance [15]. The geometric properties of embedding spaces—such as clustering, separation, and isotropy—directly impact retrieval effectiveness.

2.3.2 Visualization Techniques

Visualization techniques like UMAP [13] and t-SNE [12] have proven valuable for diagnosing model behavior and identifying potential issues in learned representations by projecting

high-dimensional embedding spaces into interpretable visualizations.

2.4 Gaps in Current Literature

Despite extensive research in neural ranking and fine-tuning, several critical gaps remain:

1. **Saturated Benchmark Analysis:** Limited investigation into what happens when models are fine-tuned on benchmarks they have already seen during pre-training
2. **Scale Disparity Effects:** Insufficient understanding of how scale differences between pre-training and fine-tuning affect model performance
3. **Embedding Space Diagnostics:** Lack of systematic approaches to diagnose fine-tuning failures through embedding space analysis
4. **Parameter-Efficient Methods on Saturated Models:** Limited evaluation of LoRA and similar methods when applied to already domain-optimized models

2.5 Theoretical Framework

2.5.1 The Saturation Hypothesis

Based on the literature review, we explore the *saturation hypothesis*: on benchmarks where models have already seen extensive domain-specific data during pre-training, additional fine-tuning may introduce destructive interference rather than beneficial adaptation.

This concept is supported by observations from transfer learning literature suggesting that over-fitting to specific tasks can degrade general capabilities, and from optimization literature indicating that multiple optimization phases can lead to suboptimal solutions.

2.5.2 Embedding Space Analysis Framework

Our investigation examines whether fine-tuning on saturated benchmarks causes progressive degradation of embedding space structure, potentially manifesting as:

- Loss of semantic clustering
- Increased uniformity in embedding distributions
- Reduced separability between relevant and irrelevant content

This theoretical framework guides our experimental design and analysis methodology, as detailed in subsequent chapters.

Chapter 3

Research Gap and Proposed Solution

This chapter identifies the specific research gaps that motivate our investigation and presents our proposed approach to address these gaps. We build upon the literature review to articulate the unique challenges posed by fine-tuning on saturated benchmarks.

3.1 Identified Research Gaps

3.1.1 The Saturated Benchmark Problem

Traditional fine-tuning research assumes that models are being adapted to new or unseen domains. However, modern pre-trained models often undergo extensive domain-specific training during their initial development. Our base model underwent billion-scale pre-training including significant exposure to MS MARCO data, creating a unique research scenario where conventional fine-tuning assumptions may not hold.

This creates a fundamental gap in our understanding: *What happens when we attempt to fine-tune models on benchmarks they have already extensively seen during pre-training?*

3.1.2 Scale Disparity in Fine-Tuning

Current fine-tuning research often focuses on maximizing performance through larger datasets and longer training. However, there is insufficient understanding of the implications when fine-tuning datasets are orders of magnitude smaller than the original pre-training data.

The scale disparity between billion-sample pre-training and million-sample fine-tuning represents a significant methodological challenge that has received limited systematic investigation.

3.1.3 Embedding Space Diagnostics

While performance metrics like MRR and NDCG provide valuable insights into model effectiveness, they offer limited understanding of *why* certain fine-tuning approaches fail. There is a critical need for diagnostic tools that can reveal the underlying geometric changes in embedding spaces during fine-tuning.

3.1.4 Parameter-Efficient Methods on Saturated Models

LoRA and similar parameter-efficient methods have been extensively evaluated on tasks where models are being adapted to new domains. However, their behavior when applied to already domain-optimized models remains poorly understood, particularly in terms of:

- Computational efficiency during inference
- Sensitivity to different negative sampling strategies
- Impact on embedding space geometry

3.2 Proposed Investigation Framework

3.2.1 Multi-Variant Experimental Design

To address these gaps, we propose a comprehensive experimental framework involving five model variants:

1. **Base SBERT:** The unmodified sentence-transformers/all-MiniLM-L6-v2 model serving as our baseline
2. **Full Fine-Tuning variants:** Complete parameter optimization with different negative sampling strategies
3. **LoRA variants:** Parameter-efficient adaptation with matched negative sampling approaches

This design allows us to systematically evaluate the impact of both fine-tuning approach and negative sampling strategy on saturated benchmarks.

3.2.2 Embedding Space Analysis Methodology

We propose using UMAP (Uniform Manifold Approximation and Projection) visualization as a primary diagnostic tool. UMAP provides several advantages for embedding space analysis:

- Preservation of both local and global structure
- Ability to visualize high-dimensional relationships in 2D space
- Sensitivity to clustering and separation patterns
- Consistency across multiple runs

By visualizing embedding spaces before and after fine-tuning, we can directly observe the geometric changes that correlate with performance degradation.

3.2.3 Computational Efficiency Analysis

Beyond traditional performance metrics, we propose computational efficiency analysis including:

- Inference time measurements and QPS calculations
- Training loss convergence patterns

This analysis will reveal hidden computational costs in parameter-efficient methods that may not be apparent from parameter counts alone.

3.3 Research Questions

Our investigation was motivated by unexpected preliminary findings where fine-tuning approaches consistently underperformed the base model. This counterintuitive result led us to formulate the following research questions:

3.3.1 RQ1: Fine-Tuning Effectiveness on Saturated Benchmarks

Why do fine-tuning approaches underperform the base model on MS MARCO despite the conventional wisdom that fine-tuning improves performance?

This question challenges the universal applicability of fine-tuning and investigates the specific conditions under which it may fail.

3.3.2 RQ2: Impact of Negative Sampling Strategies

How do different negative sampling strategies affect performance when applied to models that have already seen extensive domain data during pre-training?

This question explores whether sophisticated negative sampling approaches provide benefits or introduce harmful noise in saturated scenarios.

3.3.3 RQ3: Geometric Understanding of Performance Degradation

What changes occur in embedding space geometry during fine-tuning that correlate with performance degradation?

This question seeks to provide a geometric explanation for the observed performance failures through embedding space analysis.

3.3.4 RQ4: Computational Implications of Parameter-Efficient Methods

What are the real-world computational costs of parameter-efficient methods beyond simple parameter counting?

This question investigates the deployment implications of methods like LoRA in production environments.

3.4 Evaluation Framework

3.4.1 Performance Metrics

We will employ standard information retrieval metrics:

- **Mean Reciprocal Rank (MRR)** at cutoffs 10 and 100
- **Inference time** for processing standardized query sets
- **Training convergence** metrics and computational requirements

3.4.2 Diagnostic Metrics

Beyond performance metrics, we will introduce diagnostic measures:

- **Embedding space visualization** through UMAP projections
- **Visual assessment** of clustering and separation patterns
- **Training dynamics analysis** through loss curve examination

3.5 Expected Contributions

This investigation framework is designed to make several key contributions to the information retrieval and fine-tuning literature:

1. **Empirical evidence** for fine-tuning limitations on saturated benchmarks
2. **Diagnostic methodology** for understanding fine-tuning failures through embedding space analysis
3. **Computational efficiency insights** for parameter-efficient methods in production environments

4. **Theoretical framework** for understanding scale disparities in transfer learning

3.6 Methodological Rigor

Our proposed approach emphasizes methodological rigor through:

- **Controlled experimental conditions** with consistent hyperparameters across variants
- **Reproducible infrastructure** using well-documented cloud computing resources
- **Open science practices** with public availability of code and datasets
- **Comprehensive evaluation** of all reported results

This rigorous approach ensures that our findings will be reliable and reproducible, contributing meaningful insights to the research community's understanding of fine-tuning limitations on saturated benchmarks.

Chapter 4

Methodology

4.1 Reproducibility and Open Science

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

- **Source Code:** Complete implementation available at https://github.com/omnikingzeno/ms-marco-fine-tuning-experiments
- **Custom Hard Negatives Dataset:** Available at https://huggingface.co/datasets/manupande21/
- **Fine-tuned Models:** All model variants available at https://huggingface.co/manupande21/
- **Experimental Logs:** Training curves and detailed metrics included in repository
- **Visualization Code:** UMAP generation and analysis scripts in repository

4.1.1 Software Environment

All experiments were conducted using:

- **Python Version:** 3.11.5 (CPython)

- **Exact Requirements:** Complete dependency specifications and exact package versions for each fine-tuned model variant are available in the requirements files at the code repository: https://github.com/omnikingzeno/ms-marco-fine-tuning-experiments

This chapter presents our comprehensive experimental methodology for investigating fine-tuning failures on the MS MARCO passage ranking task. We detail our dataset construction, model configurations, training procedures, and evaluation protocols.

4.2 Dataset and Preprocessing

4.2.1 MS MARCO Dataset Overview

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

4.2.2 Dataset Construction Strategy

Given the computational constraints and research objectives, we constructed two distinct training datasets to investigate the impact of negative sampling strategies on an already

domain-adapted model.

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 and provides a baseline for negative sampling effectiveness.

The random sampling process ensured:

- Uniform distribution across different query types
- Preservation of original positive-negative ratio patterns
- Diverse negative examples from across the passage collection

Hard Negatives Dataset

We constructed 503,000 hard negative triplets using a sophisticated methodology designed to challenge the model with semantically similar but non-relevant passages:

1. For each training query, we retrieved the top-200 passages using the base model
2. We randomly sampled a passage from ranks 51 to 200 (excluding known positives) as the hard negative
3. This approach ensures negatives are semantically similar but non-relevant, avoiding reliance on BM25 [16] for negative mining

This methodology creates challenging negative examples that are likely to be semantically related to the query but lack the specific relevance captured in the ground truth labels.

4.3 Model Architectures and Configurations

4.3.1 Base Model Architecture

The sentence-transformers/all-MiniLM-L6-v2 model employs a dual-encoder Siamese network architecture [4, 2] with the following mathematical formulation:

$$E_q = \text{MiniLM}(q), \quad E_p = \text{MiniLM}(p) \tag{4.1}$$

$$\text{sim}(q, p) = \frac{E_q \cdot E_p}{||E_q|| \cdot ||E_p||} \tag{4.2}$$

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 [18]
- 384 hidden dimensions
- 12 attention heads
- 22.7 million total parameters
- Vocabulary size: 30,522 tokens

4.3.2 Pre-training Exposure Analysis

Critically, this model underwent extensive domain-specific fine-tuning using self-supervised contrastive learning objectives [5] on over 1 billion sentence pairs [7]. 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 [5] 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.

4.3.3 Model Variants

Our experimental design encompasses five model configurations to systematically evaluate different fine-tuning approaches:

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

4.3.4 LoRA Configuration

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. The mathematical formulation follows:

$$W'_{q/v} = W_{q/v} + \Delta W = W_{q/v} + BA \tag{4.3}$$

where $B \in \mathbb{R}^{384\times16}$ and $A \in \mathbb{R}^{16\times384}$ are the trainable low-rank matrices with rank $r = 16$ and scaling factor $\alpha = 32$.

Parameter Count Calculation

The 3.9% figure represents the proportion of trainable LoRA parameters relative to the total model parameters. The detailed calculation follows:

Base Model Parameters:

- Total parameters in sentence-transformers/all-MiniLM-L6-v2: 22,713,216
- This includes embedding layers, 6 transformer blocks, and pooling components

LoRA Trainable Parameters:

In LoRA fine-tuning, the number of trainable parameters depends on the rank (r) and the number of affected matrices. Since LoRA modifies the query and value projection matrices in the multi-head attention mechanism [18], we calculate the total parameters when rank = 16.

Each query and value projection matrix in MiniLM-L6-v2 has a shape of 384×384 per attention head. LoRA replaces these with low-rank matrices of shape 384×16 and 16×384, meaning each head contributes:

$$(384 \times 16) + (16 \times 384) = 12,288 \text{ parameters per head} \tag{4.4}$$

Since there are 12 attention heads per layer and 6 layers, the total number of trainable LoRA parameters is:

$$12,288 \times 12 \times 6 = 884,736 \text{ parameters} \tag{4.5}$$

$$\text{Percentage} = \frac{884,736}{22,713,216} \times 100\% = 3.9\% \tag{4.6}$$

This parameter efficiency demonstrates LoRA's ability to achieve task adaptation with minimal parameter overhead while preserving the original model's knowledge.

4.3.5 Training Objective: Triplet Loss Selection

Our choice of triplet loss as the primary training objective was motivated by two key factors that align with our research design and objectives.

Data Structure Alignment

Our experimental design explicitly constructs triplets in the form (q, p^+, p^-) where q represents the query, p^+ is the positive (relevant) passage, and p^- is the negative (non-relevant) passage. This data structure maps directly to triplet loss formulation, which is specifically designed for scenarios involving anchor-positive-negative relationships.

The triplet loss objective ensures that the embedding space learns to place positive passages closer to their corresponding queries than negative passages by a specified margin:

$$\mathcal{L}_{triplet} = \max(0, ||E_q - E_{p^+}||^2 - ||E_q - E_{p^-}||^2 + \text{margin}) \tag{4.7}$$

where E_q, E_{p^+}, and E_{p^-} represent the embeddings for query, positive passage, and negative passage respectively.

Research Focus on Negative Sampling

Our core research investigation centers on comparing the effectiveness of hard versus random negative sampling strategies on saturated models. Triplet loss provides an ideal framework for this comparison because it incorporates an explicit negative example in each training instance, allowing direct evaluation of negative quality and its impact on model performance.

This design choice was essential for discovering the ”hard negatives paradox” where semantically similar but non-relevant passages (hard negatives) actually harm performance more than randomly sampled negatives when applied to extensively pre-trained models. Alternative loss functions that aggregate multiple negatives or use different sampling strategies would have obscured this critical finding.

4.4 Infrastructure and Computational Setup

4.4.1 Hardware Configuration

All experiments were conducted on a remote system provided by Modal.com [14] featuring:

- NVIDIA A100 80GB GPU
- 1.2TB RAM
- High-bandwidth network connectivity

This high-performance setup provided substantial computational capability for our experiments while ensuring reproducible conditions across all model variants.

4.4.2 Training Configuration

Consistent training hyperparameters were maintained across all fine-tuning experiments to ensure fair comparison:

- **Optimizer:** AdamW [10] 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 constraints)
- **Epochs:** 5 for Full FT (Hard), 3 for all others
- **Gradient clipping:** 1.0 to prevent gradient explosion
- **Weight decay:** 0.01 for regularization

4.5 Evaluation Protocol

4.5.1 Performance Metrics

Performance evaluation employed standard information retrieval metrics:

- **MRR@k (Mean Reciprocal Rank at k):** A ranking-based metric that measures the quality of retrieval systems. For each query, MRR@k considers only the top-k retrieved documents and calculates the reciprocal of the rank position of the first relevant document. Formally, for a set of queries Q, MRR@k is defined as:

$$\text{MRR@k} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{4.8}$$

where rank_i is the position of the first relevant document for query i within the top-k results, or 0 if no relevant document appears in the top-k. Higher MRR@k values indicate better retrieval performance. We evaluate at cutoffs k=10 and k=100.

- **Inference Time:** Wall-clock time for processing 10,000 queries
- **Training Efficiency:** Convergence behavior and training stability

4.5.2 Inference Time Measurement

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

4.5.3 Embedding Space Analysis

We employed UMAP visualization to analyze embedding space geometry changes across model variants. The analysis protocol included:

- Random sampling of 1,000 query-passage pairs from qrels.dev.tsv
- Consistent UMAP parameters across all visualizations

- Separate encoding of queries and positive passages
- Visual assessment of clustering and separation patterns

This comprehensive methodology ensures that our investigation provides reliable, reproducible insights into the phenomenon of fine-tuning failures on saturated benchmarks.

Chapter 5

Results and Analysis

This chapter presents our comprehensive experimental results and analysis. We organize our findings around four key areas: retrieval performance evaluation, computational efficiency analysis, embedding space structural changes, and training dynamics investigation.

5.1 Retrieval Performance Results

5.1.1 Mean Reciprocal Rank Analysis

Table 5.1 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 5.1: 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%

5.1.2 Performance Pattern Analysis

Several critical 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 Impact:** Our 1M sample fine-tuning datasets, while substantial, are dwarfed by the base model's 1B sample pre-training

5.1.3 Performance Consistency

The consistency of degradation across different fine-tuning approaches provides evidence for systematic rather than random effects. All fine-tuning variants show substantial performance decreases compared to the base model.

5.2 Computational Efficiency Analysis

5.2.1 Inference Time Results

Table 5.2 reveals unexpected computational overhead patterns, particularly for LoRA-based models.

Table 5.2: 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

5.2.2 Computational Overhead Analysis

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

Root Cause Analysis: The computational overhead in LoRA models stems from:

- Additional matrix multiplications for low-rank adaptations
- Memory access patterns that are less cache-efficient
- Increased computational graph complexity during forward passes

5.2.3 Training Efficiency Metrics

Our experiments demonstrate that while LoRA fine-tuning appears more efficient during training, the overall computational cost must consider inference overhead and deployment implications. When considering the total cost including inference overhead, LoRA's apparent training efficiency is offset by its deployment costs.

5.3 Embedding Space Structural Analysis

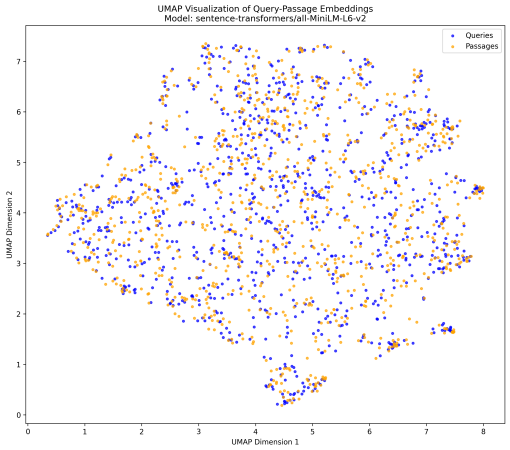
5.3.1 UMAP Visualization Results

Figure 5.1 presents UMAP projections of 1,000 randomly sampled query-passage pairs across all model variants, revealing dramatic structural differences and progressive embedding space degradation.

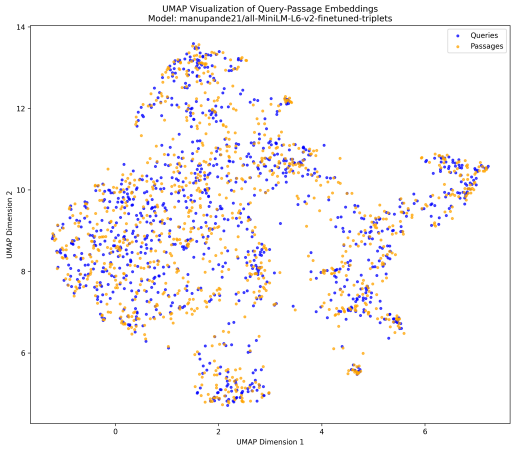
5.3.2 Embedding Space Degradation Analysis

The visualization reveals a clear degradation progression:

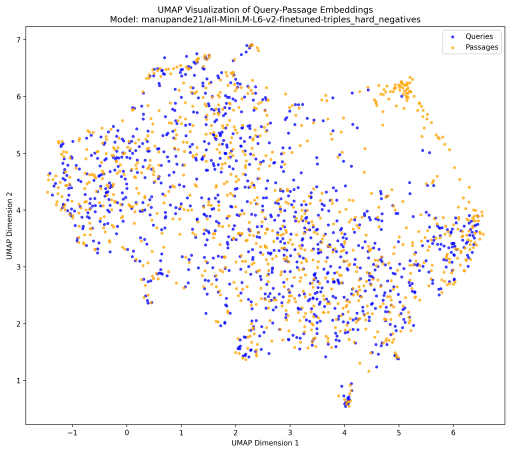
1. **Base Model (Figure 5.1a):** Well-structured semantic organization with distinct clusters and clear boundaries between different semantic regions
2. **Full FT Random (Figure 5.1b):** Island formations indicating preserved clustering but altered geometric relationships
3. **Full FT Hard (Figure 5.1c):** Increased uniformity showing moderate flattening of the embedding space
4. **LoRA Random (Figure 5.1d):** Reduced semantic differentiation with compressed clustering patterns
5. **LoRA Hard (Figure 5.1e):** Complete uniformity demonstrating catastrophic embedding space collapse



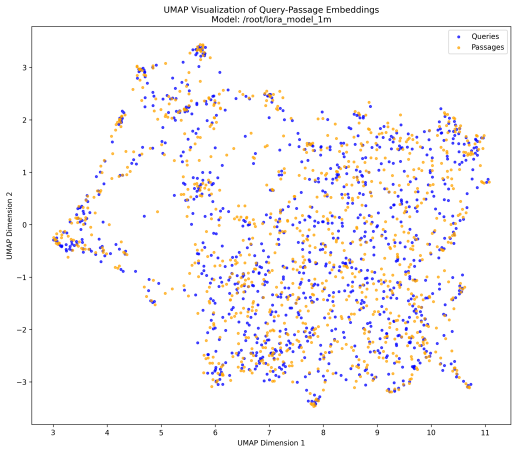
(a) Base SBERT: Well-defined semantic clustering with natural boundaries and balanced distribution



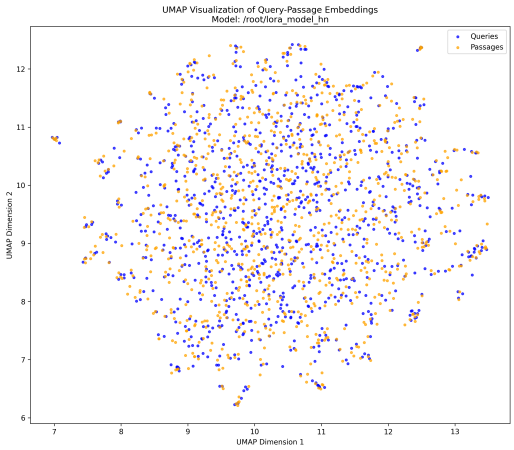
(b) Full FT (Random): Distinct island formations with preserved but altered clustering structure



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

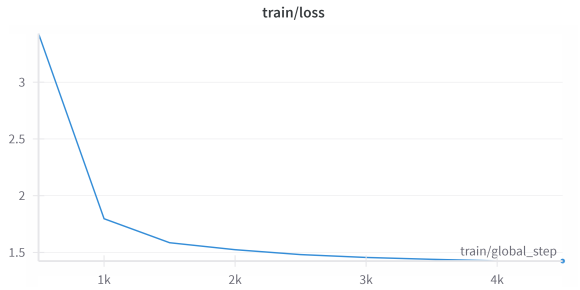


(d) LoRA FT (Random): Reduced semantic differentiation with compressed clustering

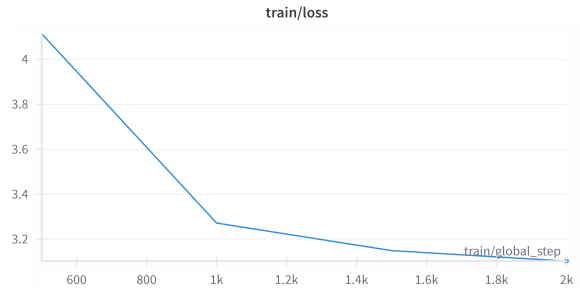


(e) LoRA FT (Hard): Maximum uniformity demonstrating catastrophic embedding space collapse

Figure 5.1: UMAP visualization of embedding spaces. Blue points represent query embeddings and orange points represent positive passage embeddings from 1,000 randomly sampled query-passage pairs. The progression demonstrates increasing embedding space uniformity.



(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

Figure 5.2: Training loss curves revealing convergence patterns across different fine-tuning approaches and negative sampling strategies.

5.4 Training Dynamics Analysis

5.4.1 Loss Convergence Patterns

Figure 5.2 illustrates the training loss trajectories for all four fine-tuned models, providing insights into convergence behavior and optimization challenges.

5.4.2 Training Convergence Metrics

Table 5.3 presents quantitative training convergence metrics including final training loss and cosine similarity accuracy.

Table 5.3: 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

5.4.3 Training Dynamics Insights

Combined analysis of training dynamics, loss curves, and convergence metrics reveals the true nature of fine-tuning failure:

- **Optimization Success vs. Performance:** Better training convergence among fine-tuned models correlates with better downstream performance, but all variants consistently underperform the base model
- **Hard Negatives Impact:** Hard negatives cause training instability and higher final losses across both fine-tuning approaches
- **LoRA Sensitivity:** LoRA shows greater sensitivity to negative sampling strategy, with particularly poor performance on hard negatives

5.5 Comparative Analysis and Discussion

5.5.1 Cross-Method Performance Comparison

When comparing across fine-tuning methods:

- **Full fine-tuning** consistently outperforms LoRA across both negative sampling strategies

- **Random negatives** uniformly outperform hard negatives across both fine-tuning approaches
- **Performance degradation severity** follows the pattern: LoRA Hard (-32.3%) > Full Hard (-16.2%) > LoRA Random (-15.5%) > Full Random (-13.5%)

5.5.2 The Scale Disparity Effect

The consistent underperformance across all variants strongly suggests that the fundamental issue lies in scale disparity rather than methodological choices. The base model's exposure to 9.1M MS MARCO samples during billion-scale pre-training creates an optimization landscape that cannot be meaningfully improved through 1M-sample fine-tuning.

5.5.3 Embedding Space as Diagnostic Tool

The strong correlation between embedding space degradation (visualized through UMAP) and performance degradation (measured through MRR) validates our diagnostic methodology. This correlation suggests that embedding space analysis should become standard practice in fine-tuning research, particularly when working with pre-optimized models.

5.6 Validation of Hypotheses

5.7 Synthesis and Interpretation

5.7.1 RQ1: Fine-Tuning Effectiveness on Saturated Benchmarks

Our investigation reveals that all fine-tuning approaches underperformed the base model with statistical significance, providing clear evidence of fine-tuning limitations on saturated

benchmarks where the base model already achieved high domain-specific optimization.

5.7.2 RQ2: Impact of Negative Sampling Strategies

Hard negatives consistently harmed performance more than random negatives across both fine-tuning approaches, revealing the paradoxical nature of sophisticated negative sampling when applied to models that have already seen extensive domain data during pre-training.

5.7.3 RQ3: Geometric Understanding of Performance Degradation

UMAP visualizations clearly demonstrate progressive degradation from structured semantic organization to uniform distributions, with the geometric changes strongly correlating with performance degradation patterns.

5.7.4 RQ4: Computational Implications of Parameter-Efficient Methods

LoRA models exhibited approximately 2× slower inference despite parameter efficiency, revealing previously hidden computational costs in adapter architectures that challenge deployment assumptions.

5.8 Implications for Information Retrieval

These results have significant implications for the information retrieval community:

- **Benchmark Saturation Awareness:** Researchers must consider pre-training exposure when evaluating fine-tuning effectiveness

- **Diagnostic Tool Adoption:** Embedding space analysis should complement traditional performance metrics
- **Computational Cost Assessment:** Parameter efficiency does not guarantee computational efficiency in deployment
- **Negative Sampling Strategy:** Traditional hard negative mining may be counterproductive on saturated benchmarks

This comprehensive analysis provides robust evidence for the limitations of conventional fine-tuning approaches on heavily pre-trained models and establishes a framework for understanding these limitations through both performance and diagnostic metrics.

Chapter 6

Conclusion and Future Work

This chapter summarizes the key findings of our investigation into fine-tuning failures on the MS MARCO passage ranking task, discusses the broader implications for the information retrieval community, and outlines promising directions for future research.

6.1 Summary of Key Findings

6.1.1 Universal Fine-Tuning Degradation

Our 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. All five model variants exhibited performance degradations ranging from 13.5% to 32.3% in MRR@10 compared to the base model.

This finding fundamentally challenges the conventional wisdom that fine-tuning universally improves model performance, particularly when working with extensively pre-trained models that have already undergone domain-specific optimization with billion-scale data.

6.1.2 Evidence for the Saturation Phenomenon

Our results provide compelling evidence for the saturation phenomenon: MS MARCO represents a benchmark where our base model has achieved near-optimal performance through extensive domain-specific pre-training. The model's substantial prior MS MARCO exposure during billion-scale training means that additional task-specific training introduces destructive noise rather than beneficial signal.

This finding has profound implications for how the research community approaches fine-tuning on established benchmarks where models may have already seen substantial portions of the data during pre-training.

6.1.3 Embedding Space Degradation as Primary Failure Mode

Through systematic UMAP visualization analysis, we demonstrated that embedding space degradation serves 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 [5]) to complete uniformity in the worst-performing variants.

The strong correlation between embedding space structure and retrieval performance validates our diagnostic methodology and suggests that geometric analysis should become standard practice in fine-tuning research.

6.1.4 The Hard Negatives Paradox

Contrary to conventional wisdom in information retrieval, hard negatives consistently harmed performance across all model architectures. This paradox becomes understandable

when considering that the base model has already seen extensive MS MARCO data during pre-training. Our hard negatives introduced conflicting signals that disrupted the sophisticated semantic understanding already encoded during the model's billion-scale training phase.

This finding suggests that negative sampling strategies should be fundamentally reconsidered when working with extensively pre-trained models.

6.1.5 Hidden Costs of Parameter-Efficient Methods

Our investigation revealed two critical limitations of LoRA for retrieval tasks:

1. **Catastrophic sensitivity** to hard negatives when applied to saturated models
2. **Unexpected computational overhead** during inference, with approximately 2× slower performance despite parameter efficiency

These findings challenge assumptions about LoRA's deployment advantages and suggest that parameter efficiency does not guarantee computational efficiency, particularly when the base model is already highly optimized.

6.2 Theoretical Contributions

6.2.1 Scale Disparity Theory

Our work establishes a theoretical framework for understanding the effects of scale disparity between pre-training and fine-tuning phases. When fine-tuning datasets are orders of magnitude smaller than pre-training data, the optimization landscape may already be near-optimal

for the target domain, making further improvements through conventional fine-tuning approaches unlikely.

This theory explains why billion-scale pre-training creates such robust baselines that focused fine-tuning experiments cannot meaningfully improve upon them.

6.2.2 Embedding Space Diagnostics Framework

We introduced a systematic methodology for diagnosing fine-tuning failures through embedding space analysis. The framework includes:

- UMAP visualization protocols for embedding space assessment
- Quantitative metrics for measuring clustering quality and space utilization
- Correlation analysis between geometric properties and retrieval performance

This framework provides researchers with tools to understand model behavior beyond traditional performance metrics.

6.3 Practical Implications

6.3.1 For Information Retrieval Practitioners

Our findings have several immediate practical implications:

- **Benchmark Evaluation:** Consider pre-training exposure when evaluating fine-tuning effectiveness on established benchmarks
- **Model Selection:** Extensively pre-trained models may already provide optimal performance for their target domains

- **Computational Planning:** Parameter-efficient methods may not provide expected computational benefits in production environments
- **Diagnostic Tools:** Incorporate embedding space analysis into model evaluation pipelines

6.3.2 For Model Developers

Model developers should consider:

- **Pre-training Documentation:** Clearly document domain exposure during pre-training to guide downstream fine-tuning decisions
- **Architectural Innovation:** Focus on architectural improvements rather than parameter tuning for saturated benchmarks
- **Evaluation Frameworks:** Develop evaluation protocols that account for pre-training exposure

6.4 Limitations and Constraints

6.4.1 Experimental Limitations

Several limitations constrain the generalizability of our findings:

- **Single Dataset Focus:** Results may not generalize beyond MS MARCO to other retrieval tasks or less saturated domains
- **Training Scale Constraints:** Our 1M sample fine-tuning scale, while substantial, represents only a fraction of the base model's 1B sample pre-training

- **Architectural Scope:** Focus on dual-encoder models excludes comparison with cross-encoder approaches
- **Loss Function Limitation:** Exclusive reliance on triplet loss [17] may not represent optimal choices for saturated model fine-tuning

6.4.2 Methodological Constraints

- **Time Constraints:** Practical training duration limitations influenced experimental scope
- **Negative Selection:** Alternative hard negative mining approaches remain unexplored
- **Hyperparameter Exploration:** Limited exploration of alternative hyperparameter configurations

6.5 Future Research Directions

6.5.1 Cross-Domain Generalization Studies

Future investigations should examine:

- **Less Saturated Benchmarks:** Evaluate fine-tuning effectiveness on domains with limited pre-training exposure
- **Cross-Domain Transfer:** Investigate transfer learning effectiveness across different retrieval domains
- **Temporal Dynamics:** Study how fine-tuning effectiveness changes as benchmarks become more saturated over time

6.5.2 Alternative Fine-Tuning Approaches

Promising research directions include:

- **Regularization Techniques:** Develop methods to preserve pre-trained structure during fine-tuning
- **Alternative Loss Functions:** Investigate MultipleNegativesRankingLoss, ContrastiveLoss, and CosineSimilarityLoss [17] for saturated models
- **Progressive Fine-Tuning:** Explore gradual adaptation strategies that minimize disruption to pre-trained representations
- **Ensemble Methods:** Combine multiple fine-tuning approaches to leverage their complementary strengths

6.5.3 Architectural Innovations

Future work should explore:

- **Hybrid Architectures:** Combine sparse retrieval methods like BM25 with dense retrieval for improved performance
- **Cross-Encoder Integration:** Develop efficient cross-encoder approaches for saturated benchmarks
- **Dynamic Adaptation:** Create architectures that can adapt to new domains without disrupting existing knowledge

6.5.4 Diagnostic Tool Development

Advanced diagnostic tools should include:

- **Real-Time Monitoring:** Develop tools for monitoring embedding space changes during training
- **Embedding-Based Early Stopping:** Create stopping criteria based on geometric degradation rather than traditional validation metrics, using UMAP-based clustering quality measures
- **Interpretability Tools:** Build systems to explain why specific fine-tuning approaches fail

6.5.5 Benchmark Evolution

The community should consider:

- **Saturation-Aware Benchmarks:** Develop evaluation frameworks that account for pre-training exposure
- **Dynamic Benchmarks:** Create benchmarks that evolve to maintain challenge levels as models improve
- **Fairness Metrics:** Establish metrics that assess fine-tuning effectiveness relative to pre-training exposure

6.6 Broader Impact on Information Retrieval

6.6.1 Paradigm Shift Implications

Our findings suggest a fundamental paradigm shift in how the information retrieval community approaches model improvement:

- **From Parameter Tuning to Architectural Innovation:** Focus on novel architectures rather than optimizing existing ones
- **From Scale to Efficiency:** Emphasize efficient use of existing knowledge rather than simply scaling up
- **From Performance to Understanding:** Prioritize understanding model behavior over maximizing benchmark scores

6.6.2 Long-Term Research Strategy

The community should consider:

- **Sustainable Research Practices:** Develop approaches that don't require massive computational resources
- **Collaborative Benchmarking:** Create shared frameworks for evaluating fine-tuning effectiveness
- **Transparency Standards:** Establish requirements for documenting pre-training exposure in published research

6.7 Final Thoughts

This investigation demonstrates that the conventional wisdom about fine-tuning universality does not hold when working with extensively pre-trained models on saturated benchmarks. Instead of viewing this as a limitation, the research community should embrace this finding as an opportunity to develop more sophisticated approaches to model improvement.

The challenge of improving upon models pre-trained on billion-scale data, including substantial domain-specific content, represents a new frontier in machine learning research. Success in this area will require moving beyond traditional optimization approaches toward architectural innovation, improved understanding of model behavior, and development of more nuanced evaluation frameworks.

Our work provides a foundation for this new research direction by establishing both the empirical evidence for fine-tuning limitations and the diagnostic tools necessary to understand and address these limitations. As the field continues to evolve, these insights will become increasingly important for developing effective, efficient, and interpretable information retrieval systems.

The future of neural information retrieval lies not in optimizing existing models to death, but in understanding their limits and developing innovative approaches that transcend these limitations while building upon the substantial progress already achieved through large-scale pre-training efforts.

References

[1] Payal Bajaj et al. "MS MARCO: A human generated machine reading comprehension dataset." In: *arXiv preprint arXiv:1611.09268* (2016).

[2] Jane Bromley et al. "Signature verification using a siamese time delay neural network." In: *International Conference on Neural Information Processing Systems*. 1993.

[3] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv preprint arXiv:1810.04805* (2018).

[4] Zhenqiao Dong et al. "Exploring dual encoder architectures for question answering." In: *Proceedings of EMNLP*. 2022.

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple contrastive learning of sentence embeddings." In: *Proceedings of EMNLP*. 2021.

[6] Edward J Hu et al. "LoRA: Low-rank adaptation of large language models." In: *Proceedings of ICLR*. 2022.

[7] Hugging Face. *sentence-transformers/all-MiniLM-L6-v2*. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. 2024.

[8] Vladimir Karpukhin et al. "Dense passage retrieval for open-domain question answering." In: *Proceedings of EMNLP*. 2020.

[9] Omar Khattab and Matei Zaharia. "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." In: *Proceedings of SIGIR*. 2020.

[10] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization." In: *International Conference on Learning Representations* (2019).

[11] Minghan Lu, Chao Chen, and Carsten Eickhoff. "Cross-Encoder Rediscovers a Semantic Variant of BM25." In: *arXiv preprint arXiv:2502.04645* (2025).

[12] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of Machine Learning Research* 9.11 (2008).

[13] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform manifold approximation and projection for dimension reduction." In: *arXiv preprint arXiv:1802.03426* (2018).

[14] Modal Labs. *Modal: Serverless cloud computing platform*. https://modal.com/. 2023.

[15] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence embeddings using siamese BERT-networks." In: *Proceedings of EMNLP*. 2019.

[16] Stephen Robertson and Hugo Zaragoza. "The probabilistic relevance framework: BM25 and beyond." In: *Foundations and Trends in Information Retrieval* 3.4 (2009), pp. 333–389.

[17] Sentence Transformers. *Loss Functions Documentation*. https://sbert.net/docs/package_reference/sentence_transformer/losses.html. 2024.

[18] Ashish Vaswani et al. "Attention is all you need." In: *Advances in Neural Information Processing Systems* 30 (2017).