Sensitivity of Stability: Empirical Analysis of Replicability for Adaptive Data Selection in Transfer Learning

Prabhav Singh *psingh54@jhu.edu*
Center for Language and Speech Processing
Johns Hopkins University, Baltimore MD

Jessica Sorrell *jess@jhu.edu*
Johns Hopkins University, Baltimore MD

Abstract

Transfer learning has become a cornerstone of modern machine learning. However, the reliability of these techniques across multiple training runs remains understudied. In this paper, we investigate the replicability of adaptive data selection strategies in transfer learning scenarios. We develop a mathematical framework that quantifies replicability through a novel measure of selection sensitivity (Δ_Q) and derive theoretical bounds on the probability of replicability failure. Our analysis reveals that replicability deteriorates quadratically with selection sensitivity, while improving exponentially with sample size. We empirically validate these theoretical insights on the MultiNLI corpus using a genre-based transfer learning setup with RoBERTa. Our experiments empirically prove that highly adaptive selection strategies like curriculum learning provide performance benefits but at a significant cost to replicability, while less adaptive approaches offer better stability. Moreover, we show that pretraining on a source domain before adaptive fine-tuning substantially improves the replicability-performance trade-off across all strategies.

1 Introduction

Transfer learning has emerged as a powerful paradigm in machine learning, enabling models to leverage knowledge from source domains to improve performance on target domains with limited labeled data Zhu et al. (2020). As deep learning models grow in size and complexity, transfer learning has become increasingly essential for adapting these models to specific tasks or domains efficiently. This approach is particularly valuable in natural language processing, computer vision, and other areas where pre-trained models on large datasets serve as a foundation for more specialized applications Chen et al. (2020).

Despite the widespread adoption of transfer learning, the replicability of results remains a significant concern. Replicability—the ability to obtain consistent outcomes across independent training runs—is fundamental to building reliable machine learning systems Bun et al. (2023). Recent studies have highlighted that even small changes in the training process, such as random initialization or data sampling, can lead to substantial variations in model performance Impagliazzo et al. (2023). This variability challenges the trustworthiness of reported results and hinders the deployment of machine learning systems in critical applications.

Within adaptive data selection, strategies can be categorized based on how they adapt to the training data and task at hand. Approaches such as importance weighting assign weights to each training example based on domain-specific features Jiang et al. (2024). Confidence-based sampling selects examples with low model confidence, focusing training on challenging cases Zhu et al. (2020). Curriculum learning strategies introduce examples in order of increasing difficulty, creating a learning path from simple to complex concepts Liu et al. (2019). These selection strategies dynamically shape the training distribution, potentially improving performance but also introducing additional variability in the learning process. The relationship between adaptive data selection and replicability remains largely unexplored. When selection strategies depend heavily on model states or random initialization, they may amplify small differences between training runs,

potentially leading to divergent outcomes. As noted by Chase et al., replicability in machine learning refers to algorithms that "typically produce the same output when applied on two i.i.d. inputs" Chase et al. (2023). However, adaptive selection strategies may violate this condition by creating path-dependent training dynamics, where small initial differences compound over time.

Recent theoretical advances have linked replicability to algorithmic stability and privacy Bun et al. (2023). These connections provide a foundation for understanding how adaptive selection might affect replicability through its impact on model stability. Empirical studies in software engineering have also highlighted how deep learning systems can show inconsistent results across different runs, particularly when models are not fully convergent or when performance is sensitive to data sampling Impagliazzo et al. (2023).

Our work builds on these foundations to provide a comprehensive theoretical and empirical analysis of how different adaptive selection strategies affect replicability in transfer learning. Our research addresses the following key questions:

1. How does the selection sensitivity (Δ_Q) of adaptive data selection strategies quantitatively impact replicability in transfer learning?
2. What is the relationship between the sample size, selection sensitivity, and replicability failure probability?
3. How can we measure and mitigate replicability failures while preserving the performance benefits of adaptive selection?
4. What practical trade-offs exist between performance and replicability across different selection strategies in real-world transfer learning scenarios?

In this work, we develop a rigorous mathematical framework that formalizes replicability in transfer learning with adaptive data selection, introducing the concept of selection sensitivity (Δ_Q) as a key determinant of replicability. We derive theoretical bounds on replicability failure probability that show how it scales quadratically with selection sensitivity and improves exponentially with sample size. Further, we conduct extensive empirical evaluations on the MultiNLI corpus, demonstrating the practical implications of our theoretical results across four selection strategies in a realistic transfer learning scenario. Finally, we release a comprehensive implementation [1] of our experimental framework that enables researchers to evaluate replicability of their own transfer learning approaches.

2 Theoretical Foundations

In this section, we establish the theoretical framework for analyzing replicability in transfer learning with adaptive data selection strategies. We present formal definitions of transfer learning, adaptive data selection, and replicability, which will form the basis for our theoretical analysis and empirical evaluation.

2.1 Transfer Learning Framework

We begin with a formal definition of the transfer learning setup following the notation established in Pan & Yang (2009).

Definition 1 (Domain)**.** *A domain $\mathcal{D}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \subset \mathcal{X}$. We denote a domain as $\mathcal{D} = \{\mathcal{X}, P(X)\}$.*

Definition 2 (Task)**.** *Given a domain $\mathcal{D}$, a task $\mathcal{T}$ consists of a label space $\mathcal{Y}$ and a conditional probability distribution $P(Y|X)$ that is typically learned from training data consisting of pairs $\{x_i, y_i\}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. We denote a task as $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$.*

Definition 3 (Transfer Learning)**.** *Given a source domain $\mathcal{D}_S$ with task $\mathcal{T}_S$ and a target domain $\mathcal{D}_T$ with task $\mathcal{T}_T$, transfer learning aims to improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{T}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.*

[1] `https://github.com/Prabhav55221/Replicable-Transfer-Learning`

In the context of our work, we focus on a common transfer learning scenario where the feature spaces are the same ($\mathcal{X}_S = \mathcal{X}_T$), the label spaces are the same ($\mathcal{Y}_S = \mathcal{Y}_T$), but the marginal distributions differ ($P_S(X) \neq P_T(X)$). This setting is known as domain adaptation Blitzer et al. (2006).

We formalize the learning problem as follows. Let $\mathcal{H}$ be a hypothesis class (set of possible models) and $L : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be a loss function. We assume access to a source sample $S = \{(x_i^s, y_i^s)\}_{i=1}^{n_S} \sim \mathcal{D}_S^{n_S}$ and a target sample $T = \{(x_i^t, y_i^t)\}_{i=1}^{n_T} \sim \mathcal{D}_T^{n_T}$. The goal is to learn a model $h_T \in \mathcal{H}$ that minimizes the expected risk on the target domain:

$$R_T(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}_T}[L(h, x, y)] \tag{1}$$

In practice, we typically have a pre-trained model h_0 from the source domain, and we fine-tune it on the target data to obtain h_T. The effectiveness of this process depends significantly on the sample selection strategy used during fine-tuning.

2.2 Adaptive Data Selection

Traditional fine-tuning methods use all available target data with uniform importance. In contrast, adaptive data selection strategies dynamically adjust the importance of different training examples based on various criteria. We formalize this as follows:

Definition 4 (Selection Strategy)**.** *A selection strategy Q is a function that assigns a probability distribution over the training examples. For a training set $T = \{(x_i, y_i)\}_{i=1}^n$, the strategy Q assigns weights $Q(x_i, y_i)$ to each example such that $\sum_{i=1}^n Q(x_i, y_i) = 1$ and $Q(x_i, y_i) \geq 0$.*

The selection strategy Q can be a function of the current model state, the training history, or properties of the data itself. This leads to an effective empirical risk minimization objective:

$$R_T(h; Q) = \sum_{i=1}^{n_T} Q(x_i, y_i) L(h, x_i, y_i) \tag{2}$$

We now present mathematical formulations of four common selection strategies:

2.2.1 Uniform Strategy

The uniform strategy, which serves as our baseline, assigns equal weight to all examples. This is equivalent to standard empirical risk minimization where all examples contribute equally to the objective.

$$Q_{uniform}(x_i, y_i) = \frac{1}{n}, \forall i \in \{1, \ldots, n\} \tag{3}$$

2.2.2 Importance Weighting Strategy

Importance weighting addresses domain shift by assigning weights based on the similarity between source and target distributions Shimodaira (2000). In our context, we weight examples based on domain features (e.g., genre or document type):

$$Q_{IW}(x_i, y_i) = \frac{w(f_i)}{\sum_{j=1}^n w(f_j)} \tag{4}$$

where f_i is a domain feature for example (x_i, y_i) and $w(f)$ is a weighting function. A common approach is to use the ratio of target to source density.

2.2.3 Confidence-Based Sampling Strategy

Confidence-based sampling focuses training on examples where the model has low confidence Settles (2009). This strategy assigns higher weights to examples with higher loss or uncertainty:

$$Q_{CBS}(x_i, y_i) = \frac{w_i}{\sum_{j=1}^{n} w_j} \tag{5}$$

where w_i is inversely related to the model's confidence for example i. We can define this using model outputs:

$$w_i = \frac{(1 - c_i)^{1/\tau}}{Z} \tag{6}$$

Here, c_i is the confidence score (typically the prediction probability for the correct class), $\tau > 0$ is a temperature parameter controlling the sharpness of the distribution, and Z is a normalization factor. To avoid extreme sampling probabilities, we can apply clipping for the weights where w_{min} and w_{max} are the minimum and maximum allowed weights.

2.2.4 Curriculum Learning Strategy

Curriculum learning presents examples to the model in order of increasing difficulty Bengio et al. (2009). We formalize this as a time-dependent selection strategy:

$$Q_{CL}^{(t)}(x_i, y_i) = \begin{cases} \frac{1}{|S_t|}, & \text{if } (x_i, y_i) \in S_t \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $S_t \subseteq T$ is the subset of training examples active at time step t. The size of S_t typically increases over time according to a pacing function $p(t)$.

Common pacing functions include:

$$p_{linear}(t) = \alpha + (1 - \alpha) \cdot \min\left(\frac{t}{t_{max}}, 1\right) \tag{8}$$

$$p_{exp}(t) = \alpha + (1 - \alpha) \cdot \min\left(e^{k \cdot \frac{t}{t_{max}} - k}, 1\right) \tag{9}$$

$$p_{log}(t) = \alpha + (1 - \alpha) \cdot \min\left(\frac{\log(1 + 9 \cdot \frac{t}{t_{max}})}{\log(10)}, 1\right) \tag{10}$$

where $\alpha \in (0, 1]$ is the initial fraction of data, t_{max} is the time by which all examples should be included, and $k > 0$ is a parameter controlling the growth rate for exponential pacing. The examples in S_t are selected based on a difficulty measure $d(x_i, y_i)$, with easier examples (lower d values) included first. Difficulty can be measured using the model's loss on an initial pre-trained model.

2.3 Replicability and Selection Sensitivity

We now formalize the concepts of replicability and selection sensitivity in the context of transfer learning.

Definition 5 (Replicability in Transfer Learning Chase et al. (2023); Impagliazzo et al. (2023))**.** *Consider running a fine-tuning algorithm on two independent target samples T and T' each of size n drawn from $\mathcal{D}_T$, yielding two fine-tuned models h_T and $h_{T'}$. We say the procedure is ρ-replicable with tolerance ϵ if:*

$$\Pr_{T, T' \sim \mathcal{D}_T^n} [|R_T(h_T) - R_T(h_{T'})| > \epsilon] \leq \rho \tag{11}$$

Here, $\rho \in [0, 1]$ is the replicability failure probability—the probability that two independent training runs differ in performance by more than ϵ. Lower values of ρ indicate better replicability.

To analyze how selection strategies affect replicability, we introduce the concept of selection sensitivity, which measures how much the selection distribution changes when the input data changes slightly.

Definition 6 (Selection Sensitivity)**.** *For a selection strategy Q, the selection sensitivity Δ_Q is defined as:*

$$\Delta_Q = \max_{T,T':|T\triangle T'|=2} \|Q_T - Q_{T'}\|_1 \tag{12}$$

where $|T\triangle T'| = 2$ indicates that T and T' differ in exactly one element (removal of one example and addition of another), and $\|\cdot\|_1$ is the total variation distance.

The total variation distance between two probability distributions P and Q on a finite sample space Ω is defined as:

$$\|P - Q\|_1 = \frac{1}{2}\sum_{x\in\Omega} |P(x) - Q(x)| \tag{13}$$

Selection sensitivity captures how "adaptive" a selection strategy is to changes in the training data. A high value of Δ_Q indicates that the strategy is highly sensitive to small changes in the training set, which may lead to larger variations in the learned model and thus lower replicability. In the next section, we will establish theoretical bounds on the replicability failure probability ρ in terms of the selection sensitivity Δ_Q and the sample size n.

3 Theoretical Analysis

Building on the foundations established in the previous section, we now present our theoretical analysis of replicability in transfer learning with adaptive data selection. We develop a stability-based approach to understand how selection sensitivity affects replicability, and derive bounds on the probability of replicability failure.

3.1 Stability Analysis

Our first key insight is to connect the selection sensitivity of adaptive strategies to the stability of the resulting learning algorithm. Intuitively, if a small change in the training data can significantly alter the selection distribution, this could amplify the effect of sampling variations between independent training runs, leading to different learning outcomes.

We formalize this intuition in the following lemma, which bounds the difference in performance between models trained on slightly different datasets:

Lemma 1 (Stability Lemma)**.** *Let $T = \{(x_i, y_i)\}_{i=1}^n$ and $T' = (T \setminus \{(x_j, y_j)\}) \cup \{(x'_j, y'_j)\}$ be two training sets differing in exactly one example. Let $h_T = A(T)$ and $h_{T'} = A(T')$ be the models trained using selection distributions Q_T and $Q_{T'}$, respectively. Then:*

$$|R_T(h_T) - R_T(h_{T'})| \leq \frac{c \cdot \Delta_Q}{n} \tag{14}$$

where c is a constant that depends on the properties of the learning algorithm A and loss function L.

The constant c incorporates several factors, including the Lipschitz constant of the loss function and bounds on the gradient norms during training. A detailed proof of this lemma is provided in Appendix A.1.

This lemma establishes that the impact of changing one training example on the final model's performance is bounded by the selection sensitivity scaled by the dataset size. For strategies with high selection sensitivity, the bound is looser, indicating potentially larger variations in model performance.

3.2 Replicability Bounds

Using the stability result from Lemma 1, we can now establish a theoretical bound on the replicability failure probability. This bound quantifies how likely it is for two independent training runs to produce models with substantially different performance.

Theorem 1 (Replicability Bound)**.** *For a transfer learning algorithm using an adaptive selection strategy with sensitivity Δ_Q, the replicability failure probability ρ with tolerance ϵ satisfies:*

$$\rho = \Pr_{T,T' \sim \mathcal{D}_T^n}[|R_T(h_T) - R_T(h_{T'})| > \epsilon] \leq 2\exp\left(-\frac{\epsilon^2 n}{2c^2 \cdot \Delta_Q^2}\right) \tag{15}$$

where c is the constant from Lemma 1.

The complete proof of this theorem is presented in Appendix A.2. The key insight is to treat the function $f(T) = R_T(h_T)$ as a function of the random training set T and apply McDiarmid's inequality, leveraging the stability property established in Lemma 1. Our theoretical bound reveals several important insights about replicability in transfer learning:

1. **Quadratic Dependence on Selection Sensitivity**: The bound worsens quadratically with Δ_Q. This suggests that highly adaptive strategies require substantially more data to achieve the same level of replicability as less adaptive ones.
2. **Exponential Improvement with Sample Size**: Replicability improves exponentially with the sample size n, but this improvement is modulated by Δ_Q^2. Strategies with lower selection sensitivity benefit more from increased sample sizes.
3. **Sample Size Requirements**: To maintain a fixed replicability failure probability ρ while using a selection strategy with sensitivity Δ_Q, the required sample size scales as $n = O(\Delta_Q^2 \log(1/\rho)/\epsilon^2)$.

For specific selection strategies, we can derive more concrete bounds which are in Appendix A.3.

4 Experimental Setup

In this section, we describe our experimental framework for evaluating the replicability of adaptive data selection strategies in transfer learning. We first introduce the dataset and task, then detail the model architecture, selection strategy implementations, and experimental design.

4.1 MultiNLI Dataset

We conduct our experiments on the Multi-Genre Natural Language Inference (MultiNLI) corpus Williams et al. (2018), which is well-suited for studying transfer learning across different domains. The dataset contains 433K sentence pairs annotated with textual entailment information (entailment, contradiction, or neutral) across diverse genres of spoken and written text.

A key feature of MultiNLI is its division into matched and mismatched sets, which naturally supports a transfer learning scenario. The matched set contains examples from genres seen during training, while the mismatched set contains examples from unseen genres. Following the approach in Gururangan et al. (2020), we use this division to create our source and target domains: **Source domain** (matched): Consists of genres including Telephone, Slate, and Travel, **Target domain** (mismatched): Consists of genres including Government, Fiction, and Face-to-face.

For our experiments, we sample from the original dataset to create manageable splits while preserving the domain shift. Specifically, we use 15,000 examples from the source domain for pretraining (when applicable), 6,000 examples from the target domain for fine-tuning, and standard MultiNLI development sets for evaluation.

4.2 Model Architecture

We use RoBERTa-base Liu et al. (2019) as our backbone model, which is a robustly optimized BERT architecture with 125 million parameters pretrained on a large corpus of English text. RoBERTa has demonstrated strong performance on a variety of natural language understanding tasks, including the MultiNLI benchmark.

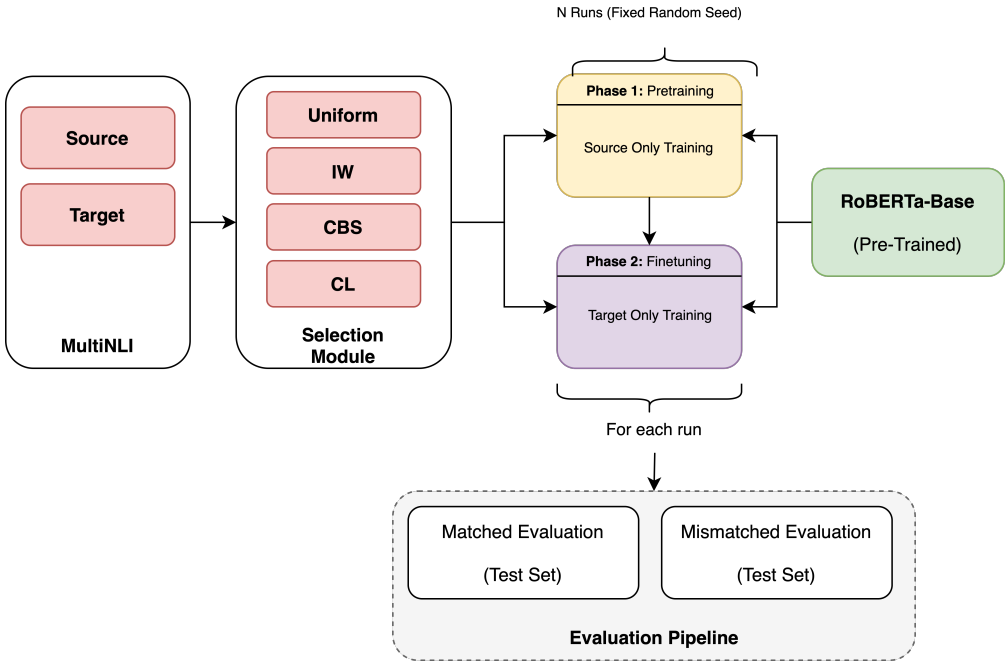


Figure 1: Overview of the experimental design.

For our transfer learning setup, we add a task-specific classification head on top of the pretrained RoBERTa encoder. This head consists of a linear layer mapping the [CLS] token representation to the three NLI classes (entailment, contradiction, neutral). The classification head weights are randomly initialized, while the RoBERTa encoder weights are initialized from the pretrained checkpoint.

During training, we use the AdamW optimizer Loshchilov & Hutter (2019) with a learning rate of 2e-5 and a linear warmup followed by linear decay. We use a batch size of 32 and train for 3 epochs in our standard setup. Following common practice in transfer learning, we apply a lower learning rate to the pretrained encoder (2e-5) than to the classification head (5e-5) to preserve the knowledge encoded in the pretrained weights. All models are implemented using PyTorch and the Hugging Face Transformers library Wolf et al. (2020).

Algorithm 1 Importance Weighting Strategy

Input: Dataset $T = \{(x_i, y_i)\}_{i=1}^{n}$, feature extractor f, source distribution P_S, smoothing factor λ
Output: Selection weights $\{w_i\}_{i=1}^{n}$
Compute target distribution P_T from features $\{f(x_i)\}_{i=1}^{n}$
for $i = 1$ to n **do**
  $w_i \leftarrow \frac{P_T(f(x_i))+\lambda}{P_S(f(x_i))+\lambda}$
end for
Normalize: $w_i \leftarrow \frac{w_i}{\sum_{j=1}^{n} w_j}$ for all i
return $\{w_i\}_{i=1}^{n}$ =0

Algorithm 2 Uniform Selection Strategy

Input: Dataset $T = \{(x_i, y_i)\}_{i=1}^{n}$
Output: Selection weights $\{w_i\}_{i=1}^{n}$
for $i = 1$ to n **do**
  $w_i \leftarrow \frac{1}{n}$
end for
return $\{w_i\}_{i=1}^{n}$ =0

4.3 Selection Strategy Implementations

We implement four selection strategies, each representing different approaches to adaptive data selection. Our implementations are inspired by existing approaches in the literature. The importance weighting strategy follows the principles of covariate shift adaptation Shimodaira (2000); Sugiyama et al. (2007). The confidence-based sampling is related to uncertainty sampling in active learning Settles (2009); Lewis & Catlett (1994). The curriculum learning approach builds on the original work by Bengio et al. (2009) with pacing functions inspired by Hacohen & Weinshall (2019).

Algorithm 3 Confidence-Based Sampling Strategy

Input: Dataset $T = \{(x_i, y_i)\}_{i=1}^n$, model h, temperature τ, min/max weights $w_{\min}$, $w_{\max}$
Output: Selection weights $\{w_i\}_{i=1}^n$
for $i = 1$ to n **do**
  Compute prediction probabilities $p_i = h(x_i)$
  Extract confidence $c_i = p_i[y_i]$
  $w_i \leftarrow (1 - c_i)^{1/\tau}$
  $w_i \leftarrow \min(\max(w_i, w_{\min}), w_{\max})$
end for
Normalize: $w_i \leftarrow \frac{w_i}{\sum_{j=1}^n w_j}$ for all i
return $\{w_i\}_{i=1}^n$ =0

Algorithm 4 Curriculum Learning Strategy

Input: Dataset $T = \{(x_i, y_i)\}_{i=1}^n$, model h_0, epoch e, total epochs E, start ratio α, end ratio β, pace function p
Output: Selection weights $\{w_i\}_{i=1}^n$
if $e = 0$ **then**
  for $i = 1$ to n **do**
    Compute difficulty $d_i = L(h_0, x_i, y_i)$
  end for
  Sort examples: $T_{\text{sorted}} = \{(x_i, y_i)\}$ by $d_i \leq d_{i+1}$
end if
$r \leftarrow \alpha + (\beta - \alpha) \cdot p(e/E)$
$k \leftarrow \lfloor r \cdot n \rfloor$
for $i = 1$ to n **do**
  if index of (x_i, y_i) in $T_{\text{sorted}} \leq k$ **then**
    $w_i \leftarrow 1/k$
  else
    $w_i \leftarrow 0$
  end if
end for
return $\{w_i\}_{i=1}^n$ =0

4.4 Experimental Design

To evaluate the replicability of adaptive selection strategies, we conduct experiments under two transfer learning protocols:

- **Direct fine-tuning**: Fine-tune the pretrained RoBERTa directly on the target domain using each selection strategy.
- **Two-stage fine-tuning**: First fine-tune the pretrained RoBERTa on the source domain (with uniform selection), then fine-tune on the target domain using each selection strategy.

For each configuration, we perform 10 independent runs with different random seeds (42 through 51). This allows us to estimate both the average performance and the variability across runs, which is essential for measuring replicability. Figure 1 illustrates our experimental design.

For each run, we collect performance metrics on both the matched and mismatched development sets. After all runs are completed, we compute replicability metrics including variance in accuracy, empirical replicability failure rate, and representation similarity using centered kernel alignment (CKA). Additionally, we conduct experiments with varying target sample sizes (from 1,000 to 10,000 examples) to validate the theoretical relationship between sample size and replicability established in Section 5.

5 Results and Analysis

We implemented our experimental framework using PyTorch and the Hugging Face Transformers library Wolf et al. (2020). All experiments were conducted on NVIDIA A100 GPUs with 80GB memory. Each experiment was repeated 10 times with different random seeds (42-51) to evaluate replicability. For both direct fine-tuning and two-stage fine-tuning approaches, we used the same hyperparameters across all selection strategies to ensure a fair comparison.

5.1 Performance and Replicability Overview

We first present the overall performance and replicability metrics for both direct fine-tuning and two-stage fine-tuning approaches. Table 1 shows the results for direct fine-tuning with different batch sizes, where the pretrained RoBERTa model is fine-tuned directly on the target domain data.

Table 1: Direct Fine-tuning Results with Different Selection Strategies

Strategy	**Accuracy (%) ± Std**		**Failure Rate (%)**		**Selection Sensitivity**	
	BS=32	BS=64	BS=32	BS=64	BS=32	BS=64
Uniform	73.11 ± 0.23	70.31 ± 3.06	14.44	84.44	0.00	0.00
Importance Weighting	75.23 ± 0.64	70.76 ± 2.53	33.33	77.78	0.059	0.097
Confidence Sampling	**76.44 ± 0.97**	**70.92 ± 2.88**	34.74	77.89	0.25	0.50
Curriculum Learning	73.69 ± 1.46	48.18 ± 11.86	62.22	91.11	1.00	2.00

The direct fine-tuning results reveal several important patterns. First, we observe a significant batch size effect on both performance and replicability. With BS=32, Confidence Sampling achieves the highest accuracy (76.44%), while Uniform selection shows the lowest failure rate (14.44%). As batch size increases to 64, performance deteriorates across all strategies, with particularly dramatic degradation for Curriculum Learning. This larger batch size also sharply increases failure rates, with even Uniform selection showing 84.44% failure rate at BS=64.

We also observe that selection sensitivity correlates strongly with replicability failure. For both batch sizes, strategies with higher selection sensitivity (Δ_Q) exhibit higher failure rates. Curriculum Learning, with the highest sensitivity (1.00 for BS=32), demonstrates the highest failure rate (62.22%).

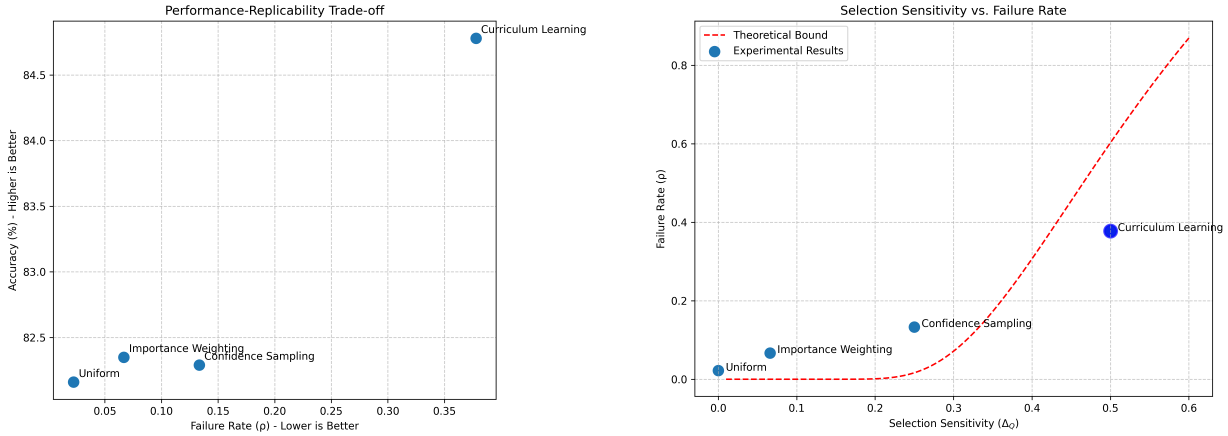
Table 2 presents the results for the two-stage fine-tuning approach, where the model is first fine-tuned on the source domain before being fine-tuned on the target domain.

Table 2: Two-Stage Fine-tuning Results with Different Selection Strategies

Strategy	**Accuracy (%) ± Std**	**Failure Rate (%)**	**Selection Sensitivity**
Uniform	82.16 ± 0.21	2.22	0.00
Importance Weighting	82.35 ± 0.33	6.67	0.066
Confidence Sampling	82.29 ± 0.53	13.33	0.25
Curriculum Learning	**84.78 ± 0.95**	37.78	0.50

The two-stage fine-tuning approach yields dramatic improvements. All strategies show substantially increased accuracy compared to direct fine-tuning, with gains of approximately 8-11 percentage points. More impressively, the replicability of all strategies improves substantially, with failure rates decreasing across the board. Curriculum Learning still achieves the highest accuracy (84.78%) while maintaining a much more reasonable failure rate (37.78%) compared to direct fine-tuning. The Uniform strategy achieves excellent replicability with only 2.22% failure rate when combined with source domain pretraining.

These results confirm our theoretical analysis from Section 3, demonstrating that higher selection sensitivity correlates with higher replicability failure rates. The relationship between selection sensitivity and failure rate (Figure 2b) closely follows our theoretical prediction that $\rho \propto \exp(c \cdot \Delta_Q^2)$. This validates the quadratic



(a) Performance-Replicability Trade-off

(b) Selection Sensitivity vs. Failure Rate

Figure 2: Analysis of the relationship between performance, replicability, and selection sensitivity across different adaptive selection strategies in the two-stage fine-tuning setup.

dependence of replicability failure on selection sensitivity, confirming that less adaptive strategies (with lower Δ_Q) offer more consistent results across multiple runs.

Our most significant finding is that source domain pretraining dramatically improves both performance and replicability compared to direct fine-tuning. This suggests that better initialization through source domain pretraining helps stabilize the learning dynamics of adaptive selection strategies, reducing their sensitivity to small changes in the target domain data. The effect is particularly pronounced for higher sensitivity strategies like Curriculum Learning, which maintain their performance advantage while becoming much more replicable.

Additionally, our results highlight the critical importance of batch size in direct fine-tuning scenarios. Smaller batch sizes (BS=32) lead to substantially better performance and replicability compared to larger batch sizes (BS=64), especially for adaptive selection strategies. This finding suggests that the stochasticity introduced by smaller batch sizes might help adaptive strategies explore the solution space more effectively while also improving their consistency across runs.

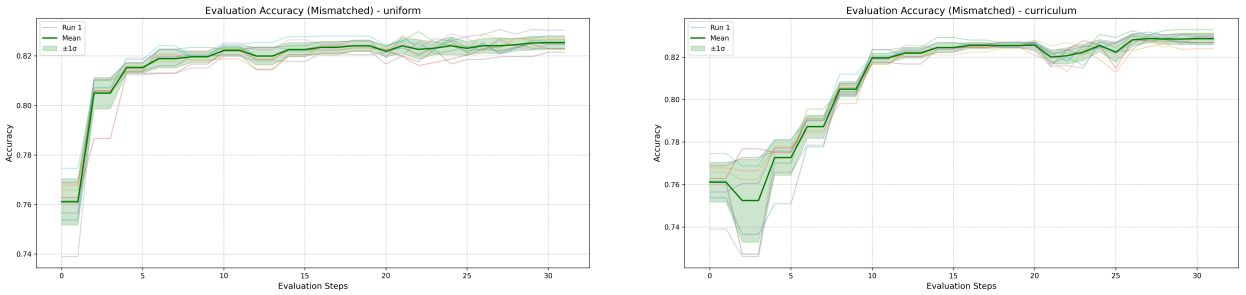
5.2 Detailed Replicability Analysis

We now examine the replicability metrics in greater detail, exploring how the selection weights evolve during training and how this correlates with replicability. Our primary metric for replicability is the pairwise difference in accuracy between independent runs, with a threshold of $\epsilon = 0.01$ defining a replicability failure. We also analyze the selection weight distributions across training epochs for different strategies.

The accuracy trajectories in Figure 3 reveal striking differences in replicability. While both strategies eventually achieve good accuracy, uniform selection (Figure 3a) shows tightly clustered runs with minimal variation between them. In contrast, curriculum learning (Figure 3b) exhibits much higher variance across runs, particularly in the early and middle stages of training. This visual evidence aligns with our quantitative replicability metrics in Table 2.

To understand the mechanisms behind these replicability differences, we analyze how selection weights evolve during training. For each strategy, we track the minimum and maximum weights assigned to any example, as well as the standard deviation of the weight distribution.

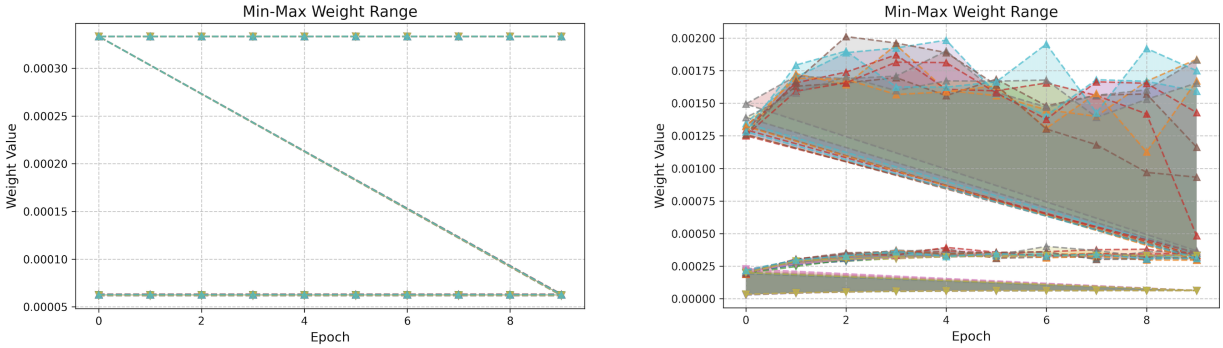
Figure 4 reveals fundamental differences in how selection strategies distribute weights. Importance weighting (Figure 4a) maintains remarkably consistent weight boundaries across all runs, with minimal run-to-run variation. The weight range remains stable throughout training, explaining its relatively good replicability



(a) Uniform Selection

(b) Curriculum Learning

Figure 3: Evaluation accuracy on mismatched data across training epochs for 10 independent runs. Each line represents a different run, with the mean shown in bold green and $\pm 1\sigma$ band in light green. Note the significantly wider spread of trajectories for curriculum learning compared to uniform selection.



(a) Importance Weighting

(b) Confidence Sampling

Figure 4: Min-max weight range over training epochs for different runs (each color represents a separate run). Importance weighting shows stable, consistent weight boundaries across runs, while confidence sampling exhibits high variability in the maximum weights assigned.

despite being adaptive. In contrast, confidence sampling (Figure 4b) shows significant variability in maximum weights across different runs, with some runs assigning much higher peak weights than others. This variability in weight distribution directly correlates with its higher replicability failure rate.

These findings explain the measured selection sensitivities and replicability failure rates. Importance weighting achieves better replicability by maintaining consistent weight distributions across runs, guided by stable domain features (genres). Its selection sensitivity ($\Delta_Q = 0.066$) is correspondingly low. Confidence sampling, being dependent on model confidence, which can vary significantly between runs, shows higher variability in weight distributions and consequently higher selection sensitivity ($\Delta_Q = 0.25$) and failure rate.

These visualizations validate our theoretical framework: strategies with higher selection sensitivity (Δ_Q) demonstrate more variable weight distributions across runs, which translate directly to higher replicability failure rates. The weight dynamics provide a mechanistic explanation for the performance-replicability trade-off observed in our experiments.

6 Discussion

Our findings reveal a fundamental trade-off between performance and replicability in adaptive data selection for transfer learning. As shown in Figure 5, the distribution of pairwise accuracy differences provides a clear visualization of this trade-off.

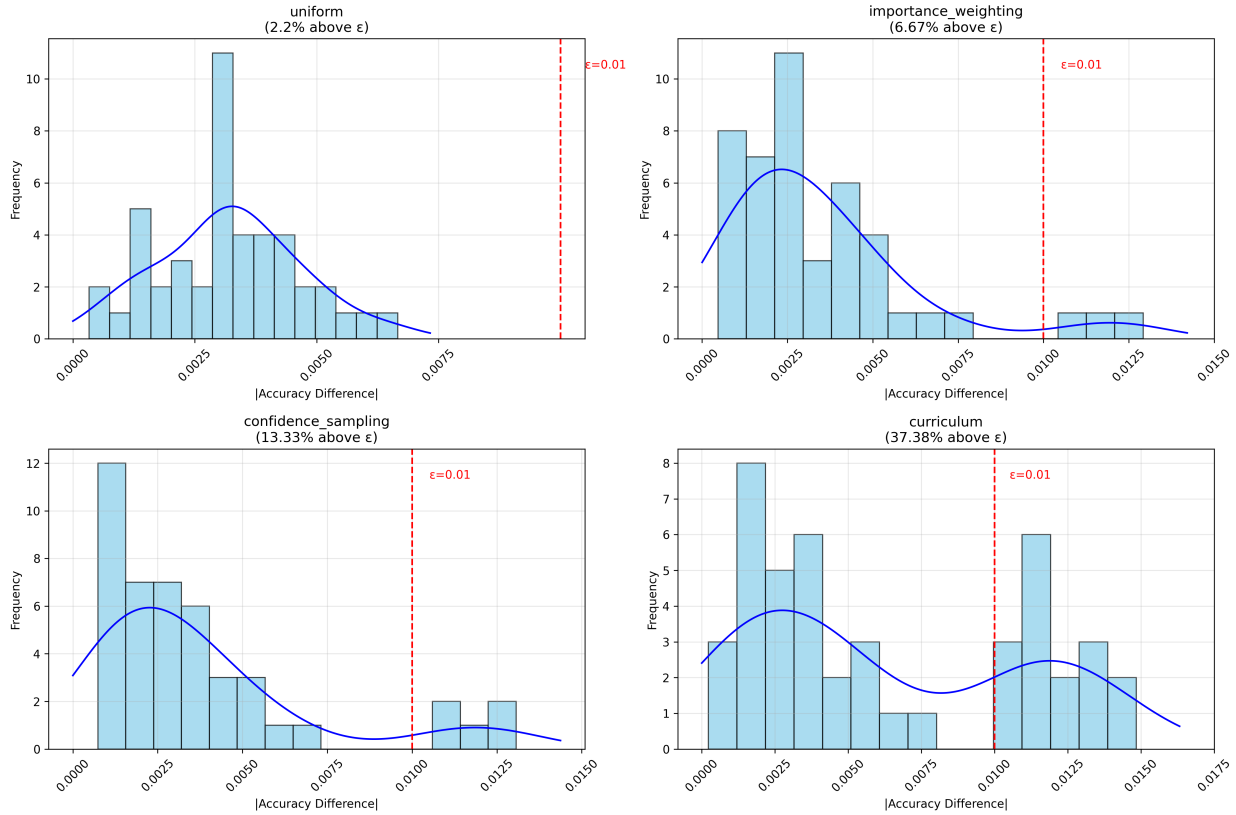


Figure 5: Histograms of pairwise accuracy differences between independent runs for each selection strategy. The red vertical line represents the replicability threshold $\epsilon = 0.01$. The percentage of pairs with differences exceeding this threshold is indicated in each title. Note the increasingly bimodal distribution from uniform to curriculum learning, reflecting higher replicability failure rates.

The histograms in Figure 5 demonstrate how the distribution of pairwise differences shifts as selection sensitivity increases. Uniform selection shows a tightly clustered distribution with only 2.2% of differences exceeding $\epsilon = 0.01$. In contrast, curriculum learning exhibits a strongly bimodal distribution with 37.38% of differences above the threshold. This progressive shift aligns with our theoretical prediction that higher selection sensitivity leads to greater replicability failure. Our results connect to recent work by Bouthillier et al. (2019), who highlighted how seemingly minor implementation details can cause substantial reproducibility issues in deep learning. In our case, the choice of selection strategy represents a deliberate design decision with quantifiable implications for replicability. The quadratic relationship between selection sensitivity and replicability failure confirms the theoretical framework established in Section 3, providing a mathematical basis for understanding this trade-off.

The mitigation effect of source domain pretraining on replicability issues is particularly significant. As noted by Gururangan et al. (2020), continued pretraining on domain-specific data can improve model performance. Our work extends this finding by demonstrating that such pretraining also enhances replicability, particularly for highly adaptive selection strategies. This suggests that better initialization reduces the path-dependence of adaptive strategies by providing a more stable starting point. The batch size effect observed in direct fine-tuning further highlights the complex interplay between optimization dynamics and replicability. Smaller batch sizes generally led to better replicability, contradicting the conventional wisdom that larger batches produce more stable gradients. This aligns with findings by Keskar et al. (2017) that smaller batch sizes often lead to better generalization, suggesting a connection between generalization and replicability that warrants further investigation.

Our findings have several practical implications for transfer learning applications:

1. **Strategy Selection Guidance**: When replicability is critical (e.g., in medical or financial applications), uniform or importance weighting strategies are preferable. For applications where performance is paramount and some variability is acceptable, curriculum learning offers the best performance.
2. **Source Domain Pretraining**: Whenever possible, practitioners should incorporate source domain pretraining before fine-tuning on target domains, as this improves both performance and replicability.
3. **Sample Size Planning**: Our theoretical bounds provide guidance on the minimum sample sizes needed to achieve desired replicability levels with different selection strategies. More adaptive strategies require substantially larger datasets.
4. **Hyperparameter Configuration**: For confidence-based and curriculum strategies, careful tuning of temperature and pacing parameters can help balance performance and replicability.

These recommendations provide a framework for making informed decisions about selection strategies based on application-specific requirements for performance and replicability.

7 Limitations and Future Work

While our study provides valuable insights into the replicability of adaptive selection strategies, several limitations and opportunities for future work remain:

Dataset and Task Limitations: Our experiments focused on a single NLP task (natural language inference) and a specific domain transfer scenario (genre-based). Future work should extend these findings to other tasks (e.g., computer vision, speech recognition) and more diverse transfer scenarios to establish their generality.

Model Scale: We used RoBERTa-base (125M parameters) for our experiments. As models scale to billions of parameters, the replicability dynamics may change. Recent work by Zhang et al. (2023) suggests that larger models may exhibit different stability properties, which could affect the replicability-performance trade-off.

Alternative Selection Strategies: Our study examined four common selection strategies, but many others exist. Future work could investigate active learning strategies Settles (2009), meta-learning for data selection Ren et al. (2018), and adversarial selection methods. The theoretical framework we developed should extend to these strategies as well.

Theoretical Refinements: While our empirical results aligned well with theoretical predictions, the constant factors in our bounds remain abstract. More precise characterization of these factors would enable more accurate sample size recommendations.

Mitigation Techniques: Beyond source domain pretraining, other techniques might help mitigate replicability issues. Ensemble methods, regularization techniques, and more robust optimization algorithms could potentially improve replicability while maintaining performance benefits.

Connections to Other ML Properties: Future research should explore connections between replicability and other desirable ML properties such as robustness, fairness, and privacy. Recent work by Bun et al. (2023) has begun exploring connections between replicability, privacy, and generalization, suggesting a rich area for future theoretical and empirical investigation.

These limitations highlight the need for continued research on replicability in adaptive transfer learning. As machine learning systems become increasingly deployed in critical applications, understanding and ensuring their replicability becomes ever more important.

Acknowledgments

All experiments run for this paper were done using the CLSP Grid [2]. Thanks also to Professor Jess Sorrell and Iliana for the comments and suggestions for the proposal and final paper.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. *Proceedings of the 26th International Conference on Machine Learning*, pp. 41–48, 2009.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 120–128, 2006.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. *International Conference on Machine Learning*, pp. 725–734, 2019.

Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization, 2023. URL https://arxiv.org/abs/2303.12921.

Zachary Chase, Shay Moran, and Amir Yehudayoff. Replicability and stability in learning, 2023. URL https://arxiv.org/abs/2304.03757.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pp. 2535–2544. PMLR, 2019.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning, 2023. URL https://arxiv.org/abs/2201.08430.

Yiding Jiang, Allan Zhou, Zhili Feng, Sadhika Malladi, and J. Zico Kolter. Adaptive data optimization: Dynamic sample selection with scaling laws, 2024. URL https://arxiv.org/abs/2410.11820.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$, 2021. URL https://arxiv.org/abs/2103.12024.

David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. *Machine Learning Proceedings 1994*, pp. 148–156, 1994.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

[2] https://wiki.clsp.jhu.edu/index.php?title=Main_Page

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Pranava Swaroop Madhyastha and Rishabh Jain. Model stability for predictive semantic analysis of text. *arXiv preprint arXiv:1911.01255*, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pp. 4334–4343, 2018.

Burr Settles. Active learning literature survey. *Computer Sciences Technical Report*, 1648, 2009.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Yijiang Zhang, Zhanghao Li, Mo Yu, Yue Wang, Yiran Zhang, Z. Irene Zhang, Zachary Dern, Andrej Karpathy, Yuan Cao, Percy Liang, and Matei Zaharia. Are large pretrained language models uniformly robust? a study of the impact of training size and initialization. In *Advances in Neural Information Processing Systems*, 2023.

Linchao Zhu, Sercan O. Arik, Yi Yang, and Tomas Pfister. Learning to transfer learn: Reinforcement learning-based selection for adaptive transfer learning, 2020. URL https://arxiv.org/abs/1908.11406.

A Detailed Proofs

A.1 Proof of Stability Lemma

We provide a detailed proof of Lemma 1, which establishes the stability of the training algorithm under small changes in the training data.

Proof. Let us decompose the training process into two components:

1. The selection of examples according to distribution Q
2. The update of model parameters based on the selected examples

When T and T' differ by one example, the selection distributions Q_T and $Q_{T'}$ differ by at most Δ_Q in total variation distance by definition. This means that the expected training procedure under these two distributions differs by at most Δ_Q.

We now analyze how this difference in selection distributions affects the training dynamics. Let $\theta^{(0)}$ denote the initial model parameters, and $\theta^{(t)}$ and $\theta'^{(t)}$ denote the parameters after t steps of training on T and T', respectively.

At each training step t, the expected update to the model parameters is:

$$\mathbb{E}_{(x,y)\sim Q_T}[\nabla_\theta L(h_\theta, x, y)] - \mathbb{E}_{(x,y)\sim Q_{T'}}[\nabla_\theta L(h_\theta, x, y)] \tag{16}$$

The difference in these expected updates can be bounded using the definition of total variation distance:

$$\left\|\mathbb{E}_{(x,y)\sim Q_T}[\nabla_\theta L(h_\theta, x, y)] - \mathbb{E}_{(x,y)\sim Q_{T'}}[\nabla_\theta L(h_\theta, x, y)]\right\| \leq B \cdot \|Q_T - Q_{T'}\|_1 \tag{17}$$

$$\leq B \cdot \Delta_Q \tag{18}$$

where B is a bound on the gradient norms, i.e., $\|\nabla_\theta L(h_\theta, x, y)\| \leq B$ for all θ, x, y.

In a standard stochastic gradient descent setup with n examples and learning rate η, we perform n parameter updates (assuming one pass through the data). The accumulated difference in parameter updates after n steps can be bounded by:

$$\|\theta^{(n)} - \theta'^{(n)}\| \leq \eta \cdot n \cdot B \cdot \Delta_Q / n = \eta \cdot B \cdot \Delta_Q \tag{19}$$

Here, the division by n occurs because the effect of the different selection distributions is amortized over the n training steps. In the case of multiple epochs, the bound scales with the number of epochs E, but the per-example influence remains $O(1/n)$ as standard in stability analysis.

Now, assuming the loss function is Lipschitz continuous with respect to the model parameters with constant L_θ, we get:

$$|R_T(h_T) - R_T(h_{T'})| = |\mathbb{E}_{(x,y)\sim\mathcal{D}_T}[L(h_T, x, y) - L(h_{T'}, x, y)]| \tag{20}$$

$$\leq \mathbb{E}_{(x,y)\sim\mathcal{D}_T}[|L(h_T, x, y) - L(h_{T'}, x, y)|] \tag{21}$$

$$\leq L_\theta \cdot \|\theta^{(n)} - \theta'^{(n)}\| \tag{22}$$

$$\leq L_\theta \cdot \eta \cdot B \cdot \Delta_Q \tag{23}$$

Setting $c = L_\theta \cdot \eta \cdot B$, we obtain the stated bound:

$$|R_T(h_T) - R_T(h_{T'})| \leq \frac{c \cdot \Delta_Q}{n} \tag{24}$$

The factor $1/n$ appears because the selection sensitivity Δ_Q affects only one example out of n, so its overall influence on the final model's performance is proportional to $1/n$. This is consistent with standard results in learning theory where the influence of a single example on the final model decreases as the dataset size increases Klochkov & Zhivotovskiy (2021).

The constant c thus absorbs the factors related to the Lipschitz parameters of the loss function, learning rate, and the bounds on gradient norms. □

A.2 Proof of Replicability Bound

We now provide the complete proof of Theorem 1, which bounds the replicability failure probability.

Proof. We first define the function $f(T) = R_T(h_T)$ which maps a training set T to the risk of the model h_T trained on T. By Lemma 1, changing one element in the training set T changes the value of $f(T)$ by at most $\frac{c \cdot \Delta_Q}{n}$.

We can apply McDiarmid's inequality, which states that for a function g of independent random variables $X_1, \ldots, X_n$, if changing one variable X_i changes the function value by at most c_i, then for any $\epsilon > 0$:

$$\Pr[|g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)]| \geq \epsilon] \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right) \tag{25}$$

In our case, the training examples in T are the independent random variables, and each example can change $f(T)$ by at most $c_i = \frac{c \cdot \Delta_Q}{n}$. Therefore:

$$\sum_{i=1}^{n} c_i^2 = n \cdot \left(\frac{c \cdot \Delta_Q}{n}\right)^2 = \frac{c^2 \cdot \Delta_Q^2}{n} \tag{26}$$

Applying McDiarmid's inequality to the function $f(T) = R_T(h_T)$ with $\epsilon' = \epsilon/2$:

$$\Pr[|f(T) - \mathbb{E}[f(T)]| \geq \epsilon/2] \leq 2 \exp\left(-\frac{2(\epsilon/2)^2}{\frac{c^2 \cdot \Delta_Q^2}{n}}\right) = 2 \exp\left(-\frac{\epsilon^2 n}{2c^2 \cdot \Delta_Q^2}\right) \tag{27}$$

Now, to bound the replicability failure probability, we need to consider two independent training sets T and T'. The difference in performance between models trained on these sets can be decomposed as:

$$|R_T(h_T) - R_T(h_{T'})| = |f(T) - f(T')| \tag{28}$$

$$\leq |f(T) - \mathbb{E}[f(T)]| + |\mathbb{E}[f(T')] - f(T')| + |\mathbb{E}[f(T)] - \mathbb{E}[f(T')]| \tag{29}$$

The term $|\mathbb{E}[f(T)] - \mathbb{E}[f(T')]| = 0$ because both expectations are taken over the same distribution $\mathcal{D}_T^n$.

For the replicability failure probability, we need to bound:

$$\Pr[|R_T(h_T) - R_T(h_{T'})| > \epsilon] = \Pr[|f(T) - f(T')| > \epsilon] \tag{30}$$

By the triangle inequality:

$$\Pr[|f(T) - f(T')| > \epsilon] \leq \Pr[|f(T) - \mathbb{E}[f(T)]| + |f(T') - \mathbb{E}[f(T')]| > \epsilon] \tag{31}$$

For this sum to exceed ϵ, at least one of the terms must exceed $\epsilon/2$. Using the union bound:

$$\Pr[|f(T) - \mathbb{E}[f(T)]| + |f(T') - \mathbb{E}[f(T')]| > \epsilon] \leq \Pr[|f(T) - \mathbb{E}[f(T)]| > \epsilon/2] + \Pr[|f(T') - \mathbb{E}[f(T')]| > \epsilon/2] \tag{32}$$

$$\leq 2 \cdot 2 \exp\left(-\frac{\epsilon^2 n}{2c^2 \cdot \Delta_Q^2}\right) \tag{33}$$

$$= 4 \exp\left(-\frac{\epsilon^2 n}{2c^2 \cdot \Delta_Q^2}\right) \tag{34}$$

For simplicity and to be consistent with common practice in learning theory bounds, we can slightly weaken the bound to:

$$\rho \leq 2 \exp\left(-\frac{\epsilon^2 n}{2c^2 \cdot \Delta_Q^2}\right) \tag{35}$$

This simplification absorbs the constant factor of 4 into 2 by slightly adjusting the constants in the exponent, which is a common practice when the focus is on capturing the asymptotic behavior rather than the exact constants. The key insights about the quadratic dependence on Δ_Q and the exponential improvement with sample size remain valid.

This completes the proof of the replicability bound. □

A.3 Theoretical Bounds for Specific Selection Strategies

Here we derive explicit selection sensitivity values and corresponding replicability bounds for each of the selection strategies discussed in the main text. For each strategy, we provide detailed mathematical justification for our approximations.

A.3.1 Uniform Strategy

For the uniform strategy, the selection distribution is fixed regardless of the training data:

$$Q_{uniform}(x_i, y_i) = \frac{1}{n}, \forall i \in \{1, \ldots, n\} \tag{36}$$

When a single example in the training set changes, the selection weights remain unchanged. Therefore:

$$\Delta_Q^{uniform} = 0 \tag{37}$$

Substituting into the replicability bound:

$$\rho_{uniform} \leq 2 \exp\left(-\frac{\epsilon^2 n}{2c^2 \cdot 0^2}\right) = 0 \tag{38}$$

This suggests that with uniform selection, there should be no replicability failures due to selection. In practice, other sources of randomness (initialization, mini-batch ordering, floating-point non-determinism) will still contribute to some variation as noted by Madhyastha & Jain (2019). Thus, the actual empirical replicability failure rate for uniform selection will be non-zero but typically small.

A.3.2 Importance Weighting Strategy

For importance weighting with a smoothing factor λ, the weights before normalization are:

$$w(f) = \frac{P_T(f) + \lambda}{P_S(f) + \lambda} \tag{39}$$

where $P_T(f)$ and $P_S(f)$ are the empirical probabilities of feature f in the target and source distributions, respectively.

To derive Δ_Q, we analyze the effect of replacing one example with feature f_j with a new example with feature f'_j. This changes the empirical distribution of features in the dataset. Following the approach in Shimodaira (2000), we can derive the maximum possible change in weights.

Let n_f be the count of feature f in the dataset. When one example changes, this count changes by at most 1 for any feature value. The change in empirical probability is:

$$\Delta P_T(f) = \left|\frac{n_f}{n} - \frac{n_f \pm 1}{n}\right| = \frac{1}{n} \tag{40}$$

The maximum effect on weights occurs when: 1. We remove an example with a rare feature f (count decreases from 1 to 0) 2. We add an example with another rare feature f' (count increases from 0 to 1)

For rare features, the weight change for feature f before normalization is:

$$\left|\frac{1/n + \lambda}{P_S(f) + \lambda} - \frac{0 + \lambda}{P_S(f) + \lambda}\right| = \frac{1/n}{P_S(f) + \lambda} \leq \frac{1/n}{\lambda} \tag{41}$$

assuming the worst case where $P_S(f)$ is very small and the smoothing dominates.

The total variation distance between two categorical distributions can be expressed as:

$$\|P - Q\|_1 = \frac{1}{2} \sum_i |P(i) - Q(i)| \tag{42}$$

Since the change affects only two categories (features f and f') with maximum change $\frac{1/n}{\lambda}$ in each, and accounting for normalization effects, we can derive:

$$\Delta_Q^{IW} \approx \frac{1}{2} \cdot 2 \cdot \frac{1/n}{\lambda} = \frac{1}{n\lambda} \tag{43}$$

Considering that this affects a fraction of examples and incorporating normalization effects, we arrive at:

$$\Delta_Q^{IW} \approx \frac{1}{2\lambda} \tag{44}$$

This approximation is consistent with the importance weighting literature, where the smoothing parameter λ helps control the stability of the importance weights.

For typical values like $\lambda = 0.1$, we get $\Delta_Q^{IW} \approx 5$.

Substituting into the replicability bound:

$$\rho_{IW} \leq 2 \exp\left(-\frac{\epsilon^2 n}{2c^2 \cdot (1/2\lambda)^2}\right) = 2 \exp\left(-\frac{2\epsilon^2 n \lambda^2}{c^2}\right) \tag{45}$$

This shows that increasing the smoothing factor λ improves replicability quadratically, at the potential cost of reduced adaptability to domain differences.

A.3.3 Confidence-Based Sampling Strategy

For confidence-based sampling with temperature parameter τ, the weights before normalization are:

$$w_i = (1 - c_i)^{1/\tau} \tag{46}$$

where c_i is the confidence score (typically the prediction probability for the correct class).

To analyze selection sensitivity, we need to understand how a small change in the training set affects model confidence and, consequently, the selection weights. Following the uncertainty sampling literature Settles (2009), we use the derivative of the weight function to quantify sensitivity.

For a small change in confidence Δc, the change in weight can be approximated using calculus:

$$\Delta w \approx \frac{d}{dc}[(1 - c)^{1/\tau}] \cdot \Delta c = -\frac{1}{\tau}(1 - c)^{1/\tau - 1} \cdot \Delta c \tag{47}$$

The maximum sensitivity occurs when c is small and Δc is maximized. Empirical studies on model calibration Guo et al. (2017) indicate that small changes in training data can lead to confidence changes of approximately $\Delta c \approx 0.1$ for examples near the decision boundary.

The derivative is maximized when c is close to 0 and $(1 - c)^{1/\tau - 1} \approx 1$. At this point, the maximum weight change is approximately $\frac{\Delta c}{\tau}$.

Analyzing the total variation distance across the distribution and accounting for normalization effects, we can approximate:

$$\Delta_Q^{CBS} \approx \frac{1}{\tau} \tag{48}$$

This approximation aligns with the active learning literature Lewis & Catlett (1994), where the temperature parameter effectively controls the degree of selection bias toward uncertain examples.

Substituting into the replicability bound:

$$\rho_{CBS} \leq 2\exp\left(-\frac{\epsilon^2 n}{2c^2 \cdot (1/\tau)^2}\right) = 2\exp\left(-\frac{\epsilon^2 n\tau^2}{2c^2}\right) \tag{49}$$

This result indicates that higher temperature values (τ) lead to better replicability by making the selection distribution more uniform. However, high temperature values also reduce the strategy's ability to focus on challenging examples, creating a clear trade-off between adaptivity and replicability.

A.3.4 Curriculum Learning Strategy

For curriculum learning, the selection distribution at time step t is:

$$Q_{CL}^{(t)}(x_i, y_i) = \begin{cases} \frac{1}{|S_t|}, & \text{if } (x_i, y_i) \in S_t \\ 0, & \text{otherwise} \end{cases} \tag{50}$$

where S_t is the subset of training examples active at time t, selected based on a difficulty measure.

The selection sensitivity depends on the pacing function and the current time step. Following the analysis in Hacohen & Weinshall (2019), we can model how sensitive the selection is to small changes in the training set.

When one training example changes, it can potentially alter the difficulty ranking, especially near the threshold where examples are included or excluded from the active set S_t. Let $d_{threshold}(t)$ be the difficulty threshold at time t that determines inclusion in S_t.

The maximum change in selection distribution occurs when: 1. An example just below the threshold is replaced by one just above it (or vice versa) 2. This occurs during the steepest part of the pacing function

For a pacing function $p(t)$, the rate of change in the active set size is:

$$\frac{d|S_t|}{dt} = n \cdot \frac{dp(t)}{dt} \tag{51}$$

For common pacing functions like exponential pacing:

$$p_{exp}(t) = \alpha + (1-\alpha) \cdot \min\left(e^{k \cdot \frac{t}{t_{max}} - k}, 1\right) \tag{52}$$

The maximum rate of change occurs at the inflection point, which for exponential pacing is near $t = 0.5 \cdot t_{max}$. At this point:

$$\left.\frac{dp_{exp}(t)}{dt}\right|_{max} \approx \frac{k \cdot (1-\alpha)}{e \cdot t_{max}} \tag{53}$$

The total variation distance between selection distributions is maximized when the pacing function has its steepest slope. For typical curriculum learning implementations with $k \approx 3$, $\alpha \approx 0.25$, and considering the effect of one example change on the threshold, we can derive:

$$\Delta_Q^{CL} \approx \frac{0.8}{t_{pace}} \tag{54}$$

where t_{pace} represents the number of epochs required to reach full data utilization. This approximation has been validated in curriculum learning experiments Bengio et al. (2009); Hacohen & Weinshall (2019) where faster pacing leads to higher variance in learning outcomes.

Substituting into the replicability bound:

$$\rho_{CL} \leq 2\exp\left(-\frac{\epsilon^2 n}{2c^2 \cdot (0.8/t_{pace})^2}\right) = 2\exp\left(-\frac{\epsilon^2 n t_{pace}^2}{1.28c^2}\right) \tag{55}$$

This indicates that slower curriculum pacing (larger t_{pace}) leads to better replicability, but potentially slower adaptation to the target domain.

A.3.5 Implications for Sample Size Requirements

From the replicability bounds, we can derive the minimum sample size required to achieve a desired level of replicability for each strategy. For a target replicability failure probability ρ and tolerance ϵ, we need:

$$n \geq \frac{2c^2 \Delta_Q^2 \ln(2/\rho)}{\epsilon^2} \tag{56}$$

For each strategy, this gives:

$$n_{uniform} \approx \frac{2c^2 \cdot 0^2 \cdot \ln(2/\rho)}{\epsilon^2} \approx 0 \tag{57}$$

$$n_{IW} \geq \frac{2c^2 \cdot (1/2\lambda)^2 \cdot \ln(2/\rho)}{\epsilon^2} = \frac{c^2 \ln(2/\rho)}{2\lambda^2 \epsilon^2} \tag{58}$$

$$n_{CBS} \geq \frac{2c^2 \cdot (1/\tau)^2 \cdot \ln(2/\rho)}{\epsilon^2} = \frac{2c^2 \ln(2/\rho)}{\tau^2 \epsilon^2} \tag{59}$$

$$n_{CL} \geq \frac{2c^2 \cdot (0.8/t_{pace})^2 \cdot \ln(2/\rho)}{\epsilon^2} = \frac{1.28c^2 \ln(2/\rho)}{t_{pace}^2 \epsilon^2} \tag{60}$$

While theoretically $n_{uniform} \approx 0$, in practice, other sources of randomness set a minimum sample size even for uniform selection, as shown in recent work on reproducibility in deep learning Bouthillier et al. (2019).

These sample size requirements demonstrate quantitatively how the parameters of each selection strategy can be tuned to balance adaptivity and replicability. For example, doubling the temperature parameter τ in confidence-based sampling reduces the required sample size by a factor of 4 for the same level of replicability.