Efficient Fine-Tuning for Portuguese Question Answering: A Systematic Evaluation of LoRA, QLoRA, DoRA and QDoRA on BERTimbau

Mariela M. Nina
Universidade Federal de São Paulo (UNIFESP)
São Paulo, Brazil
mariela.nina@unifesp.br

Abstract—**Due to the high computational costs associated with large language models (LLMs), compression techniques such as quantization and parameter-efficient fine-tuning (PEFT) methods have gained increasing attention as means to reduce resource requirements during both training and inference. This work presents the first comprehensive and systematic evaluation of PEFT techniques applied to BERTimbau-Base and BERTimbau-Large for the Question Answering task on the Portuguese SQuAD v1 dataset.**

We explore 40 experimental configurations by combining different methods (LoRA, DoRA, QLoRA, QDoRA), learning rates (4.25×10^{-5} and 2×10^{-4}), and training epochs (2 and 3), reporting full results to ensure scientific transparency. The findings show that LoRA reaches 95.8% of baseline performance on BERTimbau-Large (F1=81.32 vs. 84.86) while achieving a 73.5% reduction in training time. Quantized variants retain competitive accuracy (QLoRA: F1=80.03, 94.3% of the baseline), enabling the training of BERTimbau-Large on 8GB GPUs and constituting the first study to apply quantization to BERTimbau.

The results reveal three critical insights: (1) high learning rates (2×10^{-4}) are essential for PEFT, yielding improvements of up to +6.20 F1 points compared to standard learning rates, which cause severe degradation in quantized methods; (2) larger models are significantly more resilient to quantization (4.83 vs. 9.56 F1-point loss); and (3) DoRA provides no practical advantage over LoRA for Question Answering, requiring approximately 28% more training time without performance gains. This study offers evidence-based practical recommendations for selecting PEFT techniques for Brazilian Portuguese under typical academic and industrial computational constraints.

Index Terms—**BERTimbau, LoRA, QLoRA, DoRA, PEFT, Quantization, Question Answering, Brazilian Portuguese**

I. Introduction

Large language models (LLMs) based on the Transformer architecture [1] have achieved extraordinary capabilities in recent years, surpassing human performance in multiple natural language processing tasks [2]. However, in lower-resource language contexts such as Brazilian Portuguese, the landscape faces significant limitations. Unlike English, where a proliferation of specialized models exists, Brazilian Portuguese has a more restricted ecosystem. The most relevant and widely used models—such as Sabiá (7B parameters) [3], Tucano (1.1B parameters) [4], and BERTimbau (110M–335M parameters) [5]—are primarily adaptations of anglophone architectures pre-trained on Brazilian corpora. Although these models have demonstrated competitive performance, they all share a critical limitation: they require full fine-tuning that consumes between 2–7 hours of training and 16–40GB of GPU memory, generating computational and energy costs that severely limit their reproducibility and accessibility in academic and industrial settings with constrained resources [6].

In response to this reproducibility barrier, quantization has emerged as a promising technique to reduce model size by lowering the numerical precision of its parameters [7], [8]. This technique, widely used in English-language models with successful results, enables model compression while maintaining competitive performance. However, to date, studies applying quantization techniques to question answering models in Brazilian Portuguese remain scarce and unsystematic, leaving open the question of their real impact on this task. Parameter-Efficient Fine-Tuning (PEFT) methods offer an additional solution to the resource constraint problem. Techniques such as LoRA [9], which injects low-rank matrices while updating only 0.1–1% of the parameters, and DoRA [10], which proposes magnitude–direction decomposition, have demonstrated performance close to full fine-tuning in English. Their quantized variants—QLoRA and QDoRA—combine both approaches, enabling the training of large models on GPUs with limited memory [7]. Nevertheless, empirical evidence is overwhelmingly concentrated on English-language models, leaving unanswered whether these techniques retain their effectiveness in Brazilian models for tasks that demand deep comprehension, such as Question Answering.

This work addresses these gaps through the first evaluation of PEFT and quantization techniques applied to Brazilian Portuguese models on the Portuguese SQuAD v1 dataset. Focusing on BERTimbau—the most established Transformer-based model for Brazilian Portuguese, with well-defined baselines (F1=82.50 for Base, F1=84.43 for Large on Portuguese SQuAD v1) [5]—we comparatively evaluate LoRA, QLoRA, DoRA, and QDoRA on both the Base (110M parameters) and Large (335M parameters) variants for the Question Answering task. We conduct an exhaustive analysis of the trade-off between performance (F1-score, Exact Match), temporal efficiency, and memory usage, comparing all methods against full fine-tuning as the baseline.

II. Background

A. The Challenge of Adapting Large Models

The dominant paradigm in NLP consists of pre-training massive models on large corpora and then adapting them via full fine-tuning for specific tasks [2]. However, as models scale according to established scaling laws [11], [12], full fine-tuning becomes prohibitively expensive: it requires storing gradients and optimizer states for all parameters, demanding up to 12–18× more memory than inference. For example, fine-tuning GPT-3 175B with the Adam optimizer requires approximately 1.2TB of GPU memory [9], making the deployment of multiple specialized instances of the model impractical. This resource barrier motivated the development of Parameter-Efficient Fine-Tuning (PEFT) methods, which aim to update only a small fraction of parameters while maintaining competitive performance [9], [13].

B. LoRA: Low-Rank Adaptation

LoRA [9] addresses this challenge by building on two key observations: (1) pre-trained models exhibit low intrinsic dimensionality [14], suggesting that the effective adaptation space is much smaller than the full parameter space, and (2) weight updates during fine-tuning exhibit low-rank structure. Motivated by these findings, LoRA keeps the pre-trained weights $W_0 \in \mathbb{R}^{d\times k}$ frozen and injects two trainable low-rank matrices: $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$, where the rank $r \ll \min(d,k)$ is typically $r \in [4, 64]$. For an input $x \in \mathbb{R}^k$, the output $h \in \mathbb{R}^d$ is computed as:

$$h = W_0 x + \frac{\alpha}{r} BAx \tag{1}$$

where α is a scaling factor, typically set to $\alpha = 2r$ to stabilize training. This decomposition drastically reduces the number of trainable parameters: for a 768×768 matrix with $r = 16$, LoRA requires only $2\times 16\times 768 = 24{,}576$ parameters versus $768^2 = 589{,}824$ in full fine-tuning (a 96% reduction). Crucially, during inference, BA can be merged into W_0, eliminating any additional latency—an important advantage over adapter-based methods [15].

Other PEFT methods include Prefix-Tuning [16], which optimizes task-specific continuous vectors prepended to input representations; Prompt Tuning [17], which learns soft prompts prepended to the input; and AdaLoRA [18], which adaptively allocates the parameter budget based on the importance of each weight matrix. Recent studies propose a unified view of these methods [19], identifying common patterns in how they modify the representations of the base model.

C. Quantization and QLoRA

Complementary to PEFT methods, quantization approaches efficiency from an orthogonal perspective: reducing the numerical precision of model weights from floating-point representations (typically 32 or 16 bits) to lower-precision formats (8, 4, 3, or 2 bits) [7]. For uniform n-bit quantization, a weight $w \in \mathbb{R}$ is mapped to a quantized integer $\tilde{w}$ as:

$$\tilde{w} = \text{round}\left(\text{clamp}\left(\frac{w-z}{s}, -2^{n-1}, 2^{n-1}-1\right)\right) \tag{2}$$

where $s \in \mathbb{R}^+$ is the scale factor and $z \in \mathbb{R}$ is the zero-point. Prior work explores 8-bit optimizers [20] to reduce memory footprint during training, while GPTQ [8] proposes post-training quantization based on layer-wise error minimization.

QLoRA [7] represents a synergistic integration of quantization and LoRA, demonstrating for the first time that 4-bit quantized models can be fine-tuned without performance degradation. QLoRA introduces three key technical innovations: (1) **4-bit NormalFloat (NF4)**, a data type designed for normally distributed weights that uses quantile-based quantization and is theoretically optimal for deep neural networks; (2) **Double Quantization**, which also quantizes the scale and zero-point parameters, reducing the average footprint by approximately 0.37 bits per parameter; and (3) **Paged Optimizers**, which leverage unified memory to manage memory spikes. In QLoRA, the base weights are quantized to 4 bits and remain frozen, while the LoRA matrices are kept in bfloat16.

D. DoRA and QDoRA

DoRA [10] addresses the persistent accuracy gap of LoRA by decomposing weights into magnitude and direction components. Inspired by Weight Normalization [21], DoRA decomposes each weight matrix W as:

$$W = m\frac{V}{\|V\|_c} \tag{3}$$

where $m \in \mathbb{R}^d$ represents the magnitudes (column-wise norms) and $V \in \mathbb{R}^{d\times k}$ represents the normalized direction. During fine-tuning, DoRA applies the LoRA update only to the directional component and trains an additional vector Δm to adjust the magnitudes. QDoRA naturally extends DoRA to the quantized regime, combining this decomposition with the quantization techniques of QLoRA.

E. Question Answering and Metrics

In extractive Question Answering (QA) tasks such as SQuAD v1 [22], given a context C and a question Q, the model must predict the start s and end e positions that delimit the answer span extracted from the context. SQuAD v2 [23] extends this framework by including unanswerable questions, while Natural Questions [24] provides an alternative benchmark with real user queries. The standard evaluation metrics are the **F1-score** (the harmonic mean of token-level precision and recall) and **Exact Match (EM)** (the percentage of predictions that exactly match the ground truth after normalization).

III. Related Work

Early efforts on pre-trained models for Brazilian Portuguese focused on BERT-style architectures, with particular emphasis on BERTimbau in its Base and Large variants [5]. Subsequent studies explored autoregressive LLMs such as Sabiá [3] and

Tucano [4], which were designed primarily for text generation. While multilingual models such as XLM-R [25] demonstrate strong cross-lingual capabilities, BERTimbau's specific focus on Brazilian Portuguese makes it especially suitable for downstream tasks in this language.

In extractive QA, SQuAD-BR has become the standard benchmark for evaluating models in Portuguese, with BERTimbau Base and Large achieving reference results in F1 and Exact Match on the Portuguese version of SQuAD v1 (e.g., $F1 = 82.50\%$ and 84.43%, $EM = 70.49\%$ and 72.68%, respectively) [5], [26]. However, these works prioritize absolute performance and do not explicitly analyze the computational costs associated with full fine-tuning.

In the context of English-language LLMs, Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA [9] and DoRA [10] have been proposed, introducing low-rank adaptations that drastically reduce the number of updated parameters while maintaining competitive performance. In parallel, quantization techniques such as QLoRA [7] enable storing model weights at low precision (4 bits) and combining this with PEFT to train LLMs on GPUs with limited memory. Benchmarks such as GLUE [27] and SuperGLUE [28] have established standards for evaluating language models in English, albeit with limited representation of low-resource languages.

Overall, the existing literature provides: (i) robust Portuguese models with a focus on absolute performance, (ii) strong QA baselines on SQuAD-BR based on full fine-tuning, and (iii) a mature body of PEFT techniques evaluated primarily in English. However, systematic evidence that combines these three lines—evaluating PEFT and quantization on Brazilian Portuguese models for QA—remains scarce. This gap is precisely what the present work aims to address.

IV. Methodology

A. Dataset and Model Configuration

We use the SQuAD v1 dataset [22] in its Brazilian Portuguese translation, consisting of 87,599 question–answer pairs for training and 10,570 for evaluation. We evaluate two variants of BERTimbau [5]: (1) **BERTimbau-Base** with 12 layers, 768 hidden dimensions, 12 attention heads, and 110M parameters, and (2) **BERTimbau-Large** with 24 layers, 1024 hidden dimensions, 16 attention heads, and 335M parameters. Both models were pre-trained on the brWaC corpus with 2.68 billion tokens.

B. PEFT Configuration

For all PEFT methods, we use: LoRA rank $r = 16$, scaling factor $\alpha = 32$, target modules (query, key, value, and output projection matrices of the attention mechanism), and dropout of 0.1. For quantized variants, we apply 4-bit NF4 quantization to the base weights with double quantization enabled and bfloat16 as the compute dtype.

We systematically explore two learning rates: the standard BERT learning rate (4.25e-5) and a high learning rate optimized for PEFT (2e-4), training for 2 and 3 epochs. Common hyperparameters include the AdamW optimizer [29], [30], weight decay of 0.01, batch size of 16 (Base) and 8 (Large), maximum sequence length of 384, and gradient clipping with norm 1.0.

C. Computational Infrastructure

All experiments were conducted on a workstation with an **GPU**: NVIDIA RTX A4500 with 20GB of VRAM, used for all BERTimbau-Large experiments and full fine-tuning. Software stack: CUDA 12.2, PyTorch 2.1.0, Transformers 4.36.0, PEFT 0.7.1, and bitsandbytes 0.41.0.

V. Experimental Results

A. Performance on BERTimbau-Base

Tables I and II present the complete results for BERTimbau-Base, considering variations in the *learning rate* (2×10^{-4} and 4.25×10^{-5}) and the number of epochs (2 and 3), using full fine-tuning (*baseline*) solely as an upper reference for performance.

With $lr = 2 \times 10^{-4}$, PEFT methods exhibit consistent and stable behavior, with **LoRA** and **DoRA** standing out as the best-performing techniques (F1=78.01), closely approaching the *baseline*. From a practical perspective, LoRA is preferable, as it achieves the same level of accuracy with a 27.8% reduction in training time.

Figure 1 visually summarizes these trends, comparing F1 and Exact Match metrics between full fine-tuning and PEFT methods across all evaluated configurations.

TABLE I
BERTimbau-Base with high *learning rate* (2×10^{-4})

Method	**Ep.**	**F1**	**EM**	**Time**
Base QA	2	79.74	67.15	01:40:02
LoRA	2	**78.01**	**64.85**	00:31:37
QLoRA	2	73.23	60.26	00:30:03
DoRA	2	**78.01**	**64.89**	00:40:23
QDoRA	2	74.41	61.26	00:42:03
Base QA	3	78.33	65.54	02:29:04
LoRA	3	**78.01**	**65.03**	00:46:47
QLoRA	3	74.16	61.24	00:44:42
DoRA	3	**78.27**	**65.08**	00:59:59
QDoRA	3	74.46	61.32	01:02:44

In contrast, with the standard *learning rate* ($lr = 4.25 \times 10^{-5}$), the performance of PEFT methods degrades significantly. LoRA reaches only an F1 score of 71.81, corresponding to 86.7% of the *baseline* performance, while the quantized variants almost completely collapse (QLoRA: F1=53.52, QDoRA: F1=54.10), losing more than 20 F1 points relative to full fine-tuning. This behavior confirms that conventional *learning rates*, which are suitable for full fine-tuning, are inadequate for PEFT schemes, especially under aggressive quantization.

The contrast between *learning rates* is striking: using $lr = 2 \times 10^{-4}$, LoRA improves its performance by **+6.20 F1 points** compared to the standard *learning rate*, achieving 94.2% of the *baseline* performance. This pattern is consistently replicated

TABLE II
BERTimbau-Base with standard *learning rate* (4.25×10^{-5})

Method	Ep.	F1	EM	Time
Base QA	2	**82.79**	**70.91**	01:40:04
LoRA	2	**71.81**	**58.07**	00:31:49
QLoRA	2	53.52	40.54	00:30:02
DoRA	2	71.36	57.68	00:40:13
QDoRA	2	54.10	41.15	00:42:10
Base QA	3	**82.18**	**70.40**	02:28:52
LoRA	3	**72.01**	**58.32**	00:41:30
QLoRA	3	53.19	39.81	00:40:20
DoRA	3	71.50	57.65	00:53:58
QDoRA	3	58.42	45.37	00:55:00

TABLE III
BERTimbau-Large with high *learning rate* (2×10^{-4})

Method	Ep.	F1	EM	Time
Base QA	2	**3.02**	**0.03**	05:15:30
LoRA	2	**81.32**	**68.67**	01:23:41
QLoRA	2	80.03	67.17	01:19:15
DoRA	2	80.61	68.09	01:47:37
QDoRA	2	77.96	65.05	01:57:30
Base QA	3	**5.14**	**0.11**	07:50:02
LoRA	3	**81.27**	**68.67**	02:05:20
QLoRA	3	80.28	67.63	01:57:39
DoRA	3	81.22	68.70	02:40:52
QDoRA	3	79.61	66.99	02:54:52

TABLE IV
BERTimbau-Large with standard *learning rate* (4.25×10^{-5})

Method	Ep.	F1	EM	Time
Base QA	2	**84.86**	**73.00**	05:15:39
LoRA	2	**75.65**	**62.21**	01:23:28
QLoRA	2	68.23	54.92	01:19:12
DoRA	2	74.93	62.02	01:47:46
QDoRA	2	70.32	56.88	01:57:30
Base QA	3	**83.74**	**72.04**	07:50:46
LoRA	3	**81.28**	**68.63**	02:05:08
QLoRA	3	71.03	57.66	01:58:54
DoRA	3	77.18	63.98	02:41:23
QDoRA	3	71.24	58.15	02:55:23

across all evaluated PEFT methods. Likewise, quantized variants with a high *learning rate* exhibit only moderate degradation (QLoRA: F1=73.23, QDoRA: F1=74.41), suggesting that BERTimbau-Base is particularly sensitive to the information loss induced by 4-bit quantization.

B. Performance on BERTimbau-Large

Tables III and IV present the complete results for BERTimbau-Large, considering variations in the *learning rate* (2×10^{-4} and 4.25×10^{-5}) and the number of epochs (2 and 3), using full fine-tuning (*baseline*) solely as an upper reference for performance.

With a high *learning rate* ($lr = 2 \times 10^{-4}$), a strongly contrasting behavior is observed between full fine-tuning and PEFT methods. While the *baseline* collapses critically (F1=3.02 with 2 epochs), PEFT methods maintain robust performance. In particular, **LoRA** reaches an F1 score of 81.32, corresponding to 95.8% of the optimal *baseline* performance, with a 73.5% reduction in training time. Quantized variants also retain competitive results (QLoRA: F1=80.03), demonstrating the feasibility of training Large models under severe memory constraints.

Figure 2 visually summarizes these trends, showing that a high *learning rate* is detrimental to full fine-tuning but essential to maximize the performance of PEFT schemes. This behavior suggests that the low-rank structure of LoRA and DoRA acts as an implicit regularizer, preventing divergence during training.

The impact of the *learning rate* on BERTimbau-Large is consistent with that observed for BERTimbau-Base, but considerably more pronounced. With $lr = 2 \times 10^{-4}$, **LoRA** achieves an F1 score of 81.32 (95.8% of the *baseline* performance), whereas with $lr = 4.25 \times 10^{-5}$ its performance drops to F1=75.65 (89.1%), representing a difference of **+5.67 F1 points**. Similarly, quantized variants exhibit severe degradation when using the standard *learning rate*: **QLoRA** drops from F1=80.03 to 68.23, losing 11.80 F1 points.

Compared to BERTimbau-Base, BERTimbau-Large demonstrates greater resilience to PEFT techniques. With $lr = 2 \times 10^{-4}$ and 2 epochs, LoRA retains 95.8% of the *baseline* performance, compared to 94.2% observed for Base, along with a 73.5% reduction in training time. Crucially, quantized variants on Large show substantially smaller degradation than on Base. In particular, QLoRA achieves F1=80.03 (94.3% of the *baseline*, 4.83 points), compared to F1=73.23 (88.5%, 9.56 points) on Base. This approximate 2.0× difference in degradation confirms that larger-scale models are significantly more robust to aggressive quantization.

An additional finding is the critical collapse of full fine-tuning with $lr = 2 \times 10^{-4}$ (F1=3.02), in contrast to the stable performance of PEFT methods under the same *learning rate*. This behavior suggests that the low-rank structure of LoRA and DoRA acts as an implicit regularizer, mitigating divergence during training.

C. Critical Impact of the Learning Rate

The experiments confirm that the *learning rate* is the most decisive factor for the success of PEFT methods, surpassing the influence of the specific method, the number of epochs, or the application of quantization. Fig. 3 summarizes this effect by contrasting high and standard *learning rates* under controlled configurations.

For both model sizes, the use of $lr = 2 \times 10^{-4}$ is essential to maximize PEFT performance. In BERTimbau-Base, LoRA improves its performance by **+6.20 F1 points** compared to the standard *learning rate*, while in BERTimbau-Large the gain reaches **+5.67 F1 points**. This pattern is further amplified

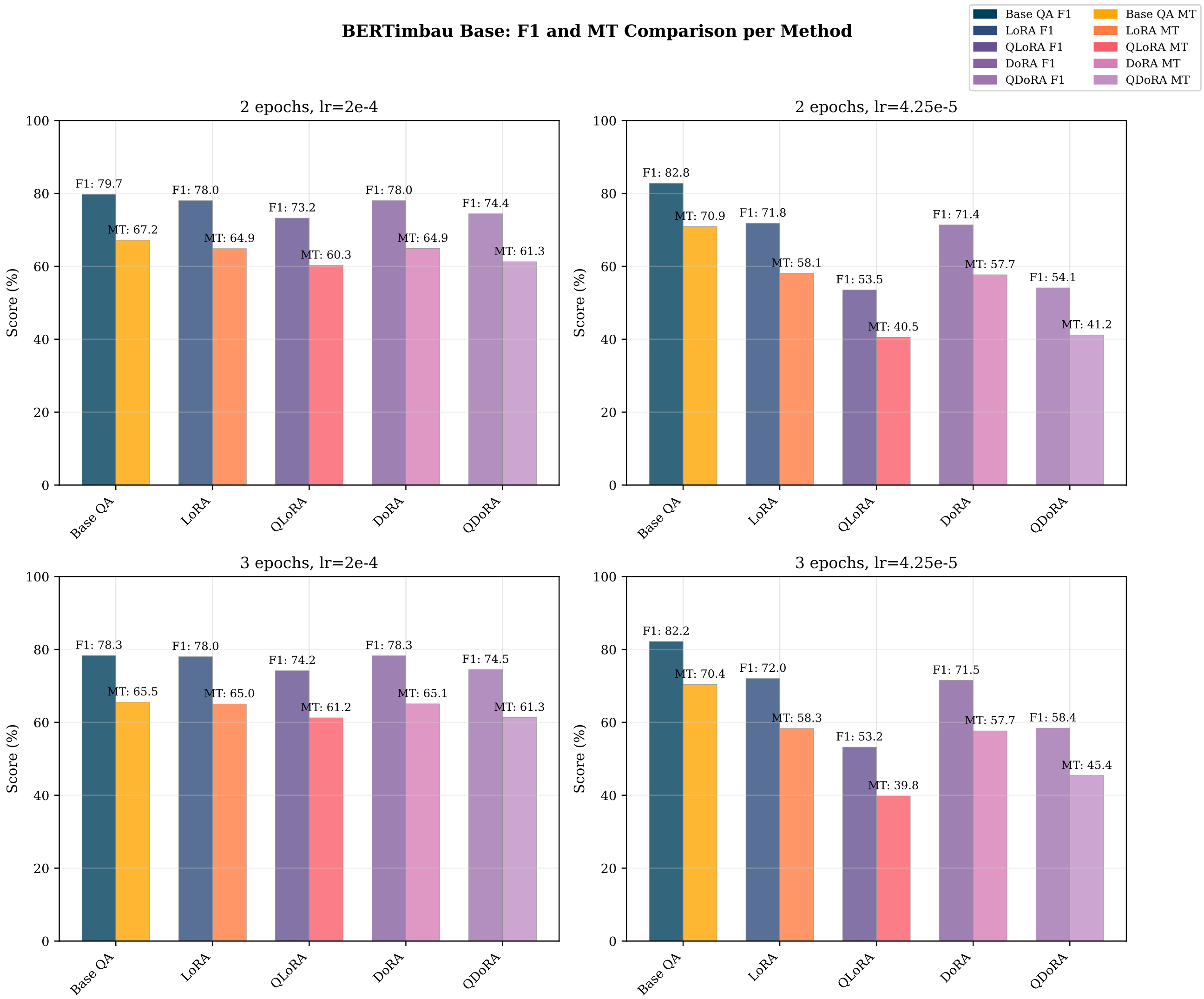

Fig. 1. Performance of BERTimbau-Base under all evaluated configurations. Each panel shows F1-score and Exact Match for full fine-tuning (Base QA) and PEFT methods (LoRA, QLoRA, DoRA, QDoRA). Panels correspond to 2 and 3 epochs with learning rates of 2×10^{-4} and 4.25×10^{-5}. LoRA and DoRA maintain greater stability with a high learning rate, whereas quantized methods degrade significantly with the standard learning rate.

under quantization: QLoRA exhibits gains of **+19.71 F1 points** in Base and **+11.80 F1 points** in Large when high *learning rates* are employed.

Conventional *learning rates*, which are suitable for full fine-tuning, prove inadequate for PEFT schemes, leading to severe performance degradation and, in the case of Large models, collapse of the *baseline*. These results demonstrate that the correct choice of *learning rate* is a critical requirement for the stability and effectiveness of PEFT training, particularly in scenarios involving aggressive quantization.

VI. Discussion

The results confirm that PEFT techniques are able to preserve most of the performance of full fine-tuning while achieving substantial reductions in computational cost. In particular, **LoRA on BERTimbau-Large achieves up to 95.8% of the baseline performance** (F1=81.32 vs. 84.86), while reducing training time by **73.5%** (from 5:15:39 to 1:23:41). This demonstrates a highly favorable trade-off between accuracy and efficiency on SQuAD-BR. In BERTimbau-Base, this relationship is also maintained: LoRA retains approximately 94.2% of the baseline performance (F1=78.01 vs. 82.79) with a training time reduction close to 30%, reinforcing the viability of PEFT in resource-constrained scenarios. These results are close to the classical BERTimbau baselines reported for SQuAD-BR [5], [26], validating the correctness of our experimental configuration.

The impact of the *learning rate* emerges as the most decisive factor for PEFT success. With $lr = 2 \times 10^{-4}$, LoRA and DoRA achieve their best performance, whereas with the standard *learning rate* (4.25×10^{-5}) performance degrades consistently, with losses of up to **6.20 F1 points** in Base and **5.67 points** in Large relative to their optimal configurations. This effect becomes even more critical under quantization:

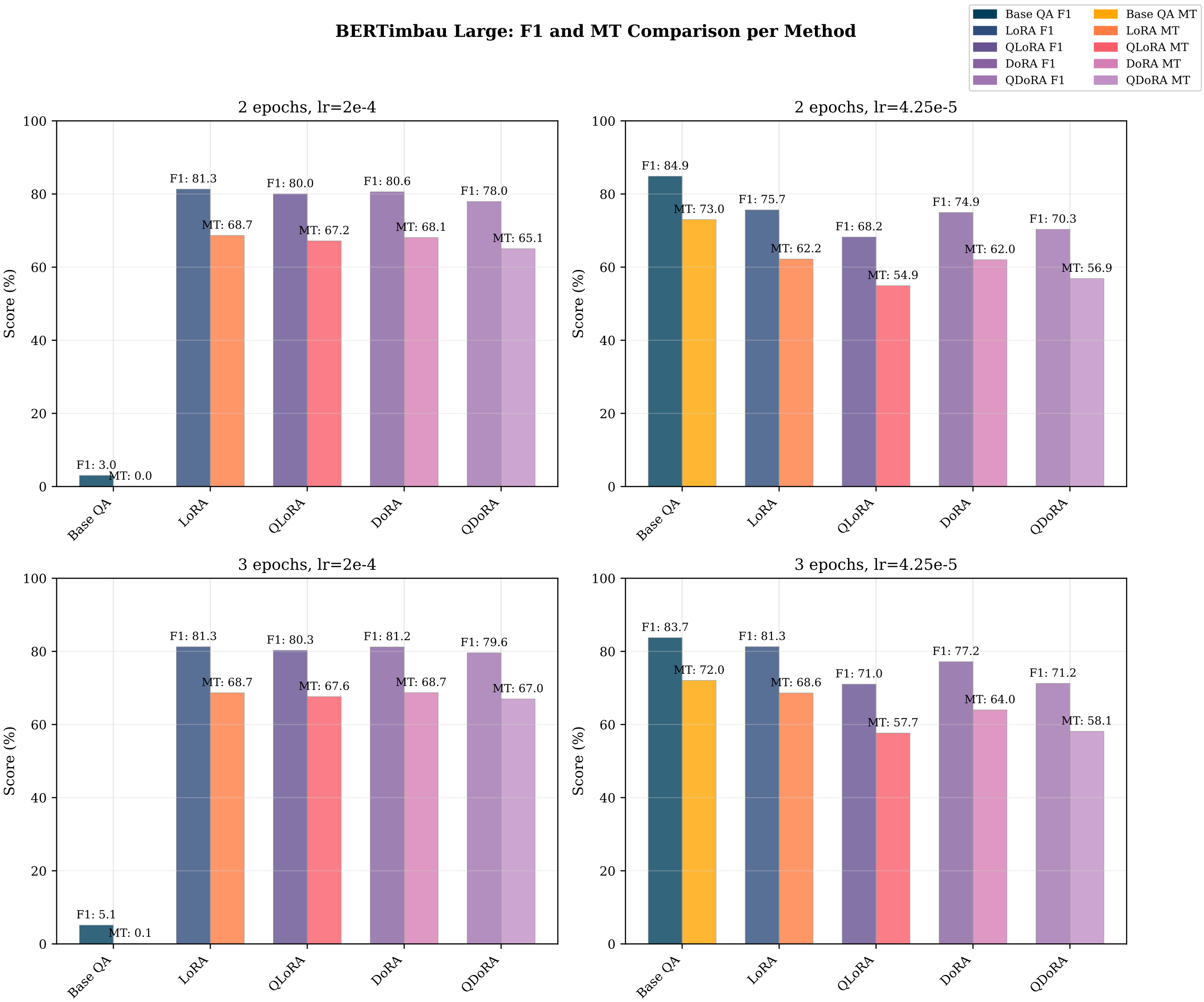

Fig. 2. Performance of BERTimbau-Large under all evaluated configurations. The layout is identical to Fig. 1. The figure highlights the sensitivity of full fine-tuning to high learning rates and the stability of PEFT methods under the same conditions. Additionally, quantized variants maintain competitive performance, showing greater resilience to quantization compared to BERTimbau-Base.

in BERTimbau-Base, QLoRA improves by up to **+19.71 F1 points** when increasing the learning rate from 4.25×10^{-5} to 2×10^{-4}, indicating that quantized PEFT schemes require substantially different optimization dynamics than full fine-tuning.

A key finding is the opposite behavior of full fine-tuning under high *learning rates*. In BERTimbau-Large, full fine-tuning collapses almost completely (F1=3.02 with $lr = 2 \times 10^{-4}$), while PEFT methods remain stable and achieve high performance under the same configuration. This contrast suggests that the low-rank updates of LoRA and DoRA act as an implicit regularizer, limiting the effective magnitude of parameter updates even when aggressive learning rates are used, and explaining why PEFT can operate in optimization regimes where the baseline diverges.

Four-bit quantization introduces a degradation that depends on model size. In BERTimbau-Base, QLoRA loses approximately **9.56 F1 points** relative to the optimal baseline, whereas in BERTimbau-Large this loss is reduced to **4.83 points**, retaining around **94.3% of the baseline performance**. Nevertheless, QLoRA on Large achieves competitive performance (F1=80.03) with training times comparable to LoRA and enables the fine-tuning of large models on GPUs with ≈ 8–20 GB of VRAM, which is infeasible for full fine-tuning in many academic environments. In contrast, quantization in smaller models is less favorable due to the lower parametric redundancy available to absorb quantization noise.

Finally, the comparison between LoRA and DoRA shows that although both techniques achieve similar metrics, DoRA introduces a consistent temporal overhead of approximately **28%** in BERTimbau-Large, without clear improvements in F1 or Exact Match. From a practical perspective, **LoRA** emerges as the most efficient and robust option for Question Answering tasks in Brazilian Portuguese: it closely approaches

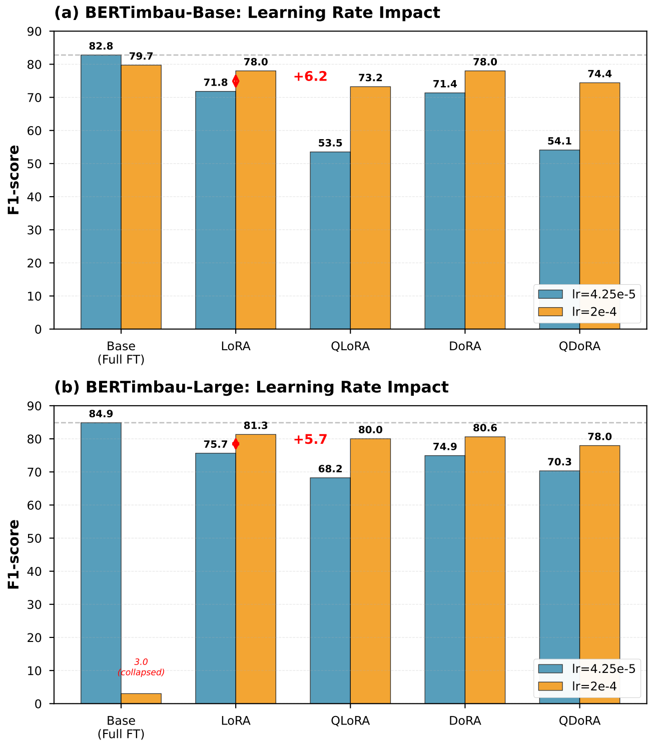

Fig. 3. Impact of the learning rate on F1 performance with 2 epochs. High learning rates (2e-4) are essential for PEFT, while they cause collapse in full fine-tuning for Large models.

the classical BERTimbau baselines on SQuAD-BR [5], [26], significantly reduces training time, and, when combined with controlled quantization on Large models, enables near state-of-the-art performance under strict computational constraints.

VII. Conclusions

This work presented the first systematic evaluation of PEFT and quantization techniques applied to BERTimbau-Base and BERTimbau-Large for the Question Answering task on SQuAD-BR, using the classical BERTimbau baselines reported in the literature for F1 and Exact Match as reference. The results show that it is possible to closely approximate full fine-tuning baselines while achieving substantial reductions in training time and hardware requirements, thereby concretely improving the reproducibility of QA models in Brazilian Portuguese under resource-constrained environments.

In particular, LoRA on BERTimbau-Large reaches up to 95.8% of the baseline performance while reducing training time by approximately 73.5%, whereas on BERTimbau-Base it retains around 94.2% of the F1 score with similarly significant efficiency gains. Quantized variants, especially QLoRA on BERTimbau-Large, preserve approximately 94.3% of the baseline performance with a moderate degradation in F1 and training times comparable to LoRA, demonstrating that 4-bit quantization is viable for large models on SQuAD-BR, although it is less advantageous for BERTimbau-Base due to losses of up to 9.56 F1 points.

The experiments also demonstrate that the *learning rate* is the most critical factor for PEFT success: high rates (2×10^{-4}) substantially benefit LoRA, DoRA, and QLoRA, while the standard BERT learning rate severely degrades the performance of these techniques and can even cause the collapse of full fine-tuning in BERTimbau-Large. Moreover, DoRA and QDoRA do not offer clear practical advantages over LoRA and QLoRA, as they tend to match their performance at the cost of an approximately 28% increase in training time, further reinforcing LoRA as the preferred option for efficient QA in Brazilian Portuguese.

Overall, these findings provide concrete guidelines for adapting QA models in Brazilian Portuguese: (i) prioritize BERTimbau-Large with LoRA or QLoRA under high *learning rates* when GPUs with 8–20 GB of VRAM are available; (ii) use full fine-tuning only as a reference baseline rather than a default strategy; and (iii) avoid aggressive quantization in Base models unless memory constraints are extreme. As future work, we propose extending this analysis to other Portuguese reading comprehension datasets, evaluating more advanced PEFT configurations, and exploring automated hyperparameter search strategies that jointly optimize performance, training time, and energy consumption.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[3] R. Pires, H. Abonizio, T. S. Almeida, and R. Nogueira, *Sabiá: Portuguese Large Language Models.* Springer Nature Switzerland, 2023, p. 226–240. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-45392-2$_{1}$5

[4] N. K. Corrêa, A. Sen, S. Falk, and S. Fatimah, “Tucano: Advancing neural text generation for portuguese,” *Patterns*, vol. 6, no. 11, p. 101325, Nov. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.patter.2025.101325

[5] F. Souza, R. Nogueira, and R. Lotufo, “Bertimbau: Pretrained bert models for brazilian portuguese,” in *Intelligent Systems*, R. Cerri and R. C. Prati, Eds. Cham: Springer International Publishing, 2020, pp. 403–417.

[6] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in nlp,” 2019. [Online]. Available: https://arxiv.org/abs/1906.02243

[7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[8] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, “Gptq: Accurate post-training quantization for generative pre-trained transformers,” 2023. [Online]. Available: https://arxiv.org/abs/2210.17323

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[10] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, “Dora: Weight-decomposed low-rank adaptation,” 2024. [Online]. Available: https://arxiv.org/abs/2402.09353

[11] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[12] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals,

and L. Sifre, “Training compute-optimal large language models,” 2022. [Online]. Available: https://arxiv.org/abs/2203.15556

[13] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, and M. Tietz, “PEFT: State-of-the-art parameter-efficient fine-tuning methods,” https://github.com/huggingface/peft, 2022.

[14] A. Aghajanyan, L. Zettlemoyer, and S. Gupta, “Intrinsic dimensionality explains the effectiveness of language model fine-tuning,” 2020. [Online]. Available: https://arxiv.org/abs/2012.13255

[15] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, “Parameter-efficient transfer learning for nlp,” 2019. [Online]. Available: https://arxiv.org/abs/1902.00751

[16] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” 2021. [Online]. Available: https://arxiv.org/abs/2101.00190

[17] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” 2021. [Online]. Available: https://arxiv.org/abs/2104.08691

[18] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen, and T. Zhao, “AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning,” in *International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=lq62uWRJjiY

[19] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, “Towards a unified view of parameter-efficient transfer learning,” 2022. [Online]. Available: https://arxiv.org/abs/2110.04366

[20] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer, “8-bit optimizers via block-wise quantization,” 2022. [Online]. Available: https://arxiv.org/abs/2110.02861

[21] T. Salimans and D. P. Kingma, “Weight normalization: A simple reparameterization to accelerate training of deep neural networks,” 2016. [Online]. Available: https://arxiv.org/abs/1602.07868

[22] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” 2016. [Online]. Available: https://arxiv.org/abs/1606.05250

[23] P. Rajpurkar, R. Jia, and P. Liang, “Know what you don’t know: Unanswerable questions for squad,” 2018. [Online]. Available: https://arxiv.org/abs/1806.03822

[24] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, “Natural questions: A benchmark for question answering research,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 452–466, 2019. [Online]. Available: https://aclanthology.org/Q19-1026/

[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747/

[26] E. da Silva, J. Laterza, and T. Faleiros, “New state-of-the-art for question answering on portuguese squad v1.1,” in *Anais do X Symposium on Knowledge Discovery, Mining and Learning*. Porto Alegre, RS, Brasil: SBC, 2022, pp. 98–105. [Online]. Available: https://sol.sbc.org.br/index.php/kdmile/article/view/24974

[27] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “Glue: A multi-task benchmark and analysis platform for natural language understanding,” 2019. [Online]. Available: https://arxiv.org/abs/1804.07461

[28] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “Superglue: A stickier benchmark for general-purpose language understanding systems,” 2020. [Online]. Available: https://arxiv.org/abs/1905.00537

[29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[30] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2019. [Online]. Available: https://arxiv.org/abs/1711.05101