Efficient Fine-Tuning Methods for Portuguese Question Answering: Evaluating LoRA, QLoRA, DoRA and QDoRA on BERTimbau with SQuAD

Mariela M. Nina
Federal University of 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 popularity to reduce resources during inference and training. This work presents the first comprehensive benchmark of PEFT techniques applied to BERTimbau-Base and BERTimbau-Large for the Question Answering task on the Portuguese SQuAD v1 dataset, executed on standard research hardware (GPUs with 8–20GB VRAM). We systematically explore 40 configurations combining methods (LoRA, DoRA, QLoRA, QDoRA), learning rates (4.25e-5 and 2e-4), and epochs (2 and 3), reporting complete results for scientific transparency. Findings demonstrate that LoRA reaches 95.8% of baseline performance in BERTimbau-Large (F1=81.32 vs. 84.86) with 73.5% training time reduction. Quantized variants maintain competitive accuracy (QLoRA: F1=80.03, 94.3% of baseline), enabling BERTimbau-Large training on 8GB GPUs, marking the first quantization study applied to BERTimbau. Results reveal three critical findings: (1) high learning rates (2e-4) are essential for PEFT, improving +6.20 F1 points over standard learning rates that collapse quantized methods, (2) larger models are significantly more resilient to quantization (4.83 vs. 9.56 F1 points loss), and (3) DoRA offers no practical advantage over LoRA in QA, requiring +28% more time without performance gains. This study provides evidence-based practical recommendations for selecting PEFT techniques in Brazilian Portuguese under typical computational constraints of academic and industrial environments.**

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 languages such as Brazilian Portuguese, there are significant limitations. Unlike English, with a proliferation of specialized models, 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 mainly adaptations of pre-trained English architectures on Brazilian corpora. Although these models show competitive performance, they share a critical limitation: full fine-tuning requires 2–7 hours of training and 16–40GB of GPU memory, generating computational and energy costs that severely limit reproducibility and accessibility in academic and industrial contexts with limited resources [6].

Quantization emerges as a promising technique to reduce model size by lowering the numerical precision of its parameters [7], [8]. Widely used in English-language models with successful results, it allows compression while maintaining competitive performance. However, quantization has not yet been applied to Brazilian Portuguese models, representing a critical gap in the literature. Complementarily, Parameter-Efficient Fine-Tuning (PEFT) methods provide another resource-saving solution. Techniques such as LoRA [9], which injects low-rank matrices updating only 0.1–1% of parameters, and DoRA [10], which decomposes magnitude-direction, have shown near full fine-tuning performance in English. Their quantized variants—QLoRA and QDoRA—combine both approaches, enabling large model training on limited-memory GPUs [7]. However, empirical evidence is overwhelmingly focused on English, leaving open whether these methods are effective for Brazilian Portuguese models in tasks demanding deep understanding such as Question Answering.

This work addresses these gaps through the first evaluation of PEFT and quantization applied to Brazilian Portuguese models. Focusing on BERTimbau—the most established transformer for Brazilian Portuguese with established baselines (F1=82.50 Base, F1=84.43 Large for Portuguese SQuAD v1) [5]—we comparatively evaluate LoRA, QLoRA, DoRA, and QDoRA in Base (110M parameters) and Large (335M parameters) variants for Question Answering. We systematically analyze the trade-off between performance (F1-score, Exact Match), time efficiency, and memory usage, comparing against full fine-tuning as baseline.

II. Background

A. The Challenge of Large Model Adaptation

The dominant NLP paradigm pre-trains massive models on large corpora and then adapts them through full fine-tuning for specific tasks [2]. However, as models grow according to 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

Adam requires approximately 1.2TB GPU memory [9], making deployment of multiple specialized instances impractical. This resource barrier motivated Parameter-Efficient Fine-Tuning (PEFT) methods, which update only a small fraction of parameters while maintaining competitive performance [9], [13].

B. LoRA: Low-Rank Adaptation

LoRA [9] addresses this challenge based on two key observations: (1) pre-trained models have a low intrinsic dimension [14], suggesting the effective adaptation space is much smaller than the full parameter space, and (2) weight updates during fine-tuning exhibit low-rank structure. LoRA keeps pretrained 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 $B \in \mathbb{R}^{d \times r}$, with rank $r \ll \min(d, k)$, typically $r \in [4, 64]$. For input $x \in \mathbb{R}^k$, 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 as $\alpha = 2r$. This decomposition drastically reduces trainable parameters: for a 768×768 matrix with $r = 16$, LoRA requires only $2 \cdot 16 \cdot 768 = 24,576$ parameters versus $768^2 = 589,824$ in full fine-tuning (96% reduction). During inference, BA can be merged with W_0, eliminating extra latency, an advantage over adapter-based methods [15].

Other PEFT methods include Prefix-Tuning [16], optimizing task-specific continuous vectors prepended to input representations; Prompt Tuning [17], learning soft prompts prepended to inputs; and AdaLoRA [18], adaptively allocating parameter budgets. Recent work provides a unified view of these methods [19].

C. Quantization and QLoRA

Quantization reduces model weight precision from floating-point (32/16 bits) to lower precision (8, 4, 3, or 2 bits) [7]. For n-bit uniform quantization, weight $w \in \mathbb{R}$ maps to integer $\tilde{w}$:

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

where $s > 0$ is the scale and z is zero-point. Prior work explores 8-bit optimizers [20] to reduce memory footprint, while GPTQ [8] proposes post-training quantization via layerwise error minimization.

QLoRA [7] synergistically combines quantization and LoRA, showing 4-bit quantized models can be fine-tuned without performance degradation. Innovations include (1) **4-bit NormalFloat (NF4)** optimized for normally distributed weights, (2) **Double Quantization**, also quantizing scale and zero-point, reducing average footprint by 0.37 bits per parameter, and (3) **Paged Optimizers**, using unified memory for peak management. Base weights are frozen at 4-bit, LoRA matrices remain in bfloat16.

D. DoRA and QDoRA

DoRA [10] addresses residual LoRA precision gaps via magnitude-direction decomposition:

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

where $m \in \mathbb{R}^d$ contains magnitudes (column norms) and $V \in \mathbb{R}^{d \times k}$ the normalized direction. During fine-tuning, LoRA updates apply only to directional component; Δm adjusts magnitudes. QDoRA extends DoRA to quantized regimes.

E. Question Answering and Metrics

In extractive QA (SQuAD v1) [21], given context C and question Q, models predict start s and end e positions for the answer span. SQuAD v2 [22] adds unanswerable questions; Natural Questions [23] provides a real-user benchmark. Standard metrics are **F1-score** (token-level harmonic mean of precision and recall) and **Exact Match (EM)**.

III. Related Work

Early pretrained models for Brazilian Portuguese focused on BERT architectures, notably BERTimbau Base and Large [5]. Later studies explored autoregressive LLMs like Sabiá [3] and Tucano [4] for text generation. Multilingual models like XLM-R [24] show cross-lingual abilities, but BERTimbau's focus on Brazilian Portuguese makes it well-suited for downstream tasks.

In extractive QA, SQuAD-BR is the standard benchmark, with state-of-the-art F1 and EM scores on SQuAD v1. Prior Portuguese NLP work includes NER [25] and other tasks, but computational costs were not addressed.

For English LLMs, PEFT methods such as LoRA [9] and DoRA [10] reduce updated parameters while maintaining performance. Quantization (QLoRA [7]) allows low-precision weight storage (4-bit) combined with PEFT for training on limited-memory GPUs. Benchmarks like GLUE [26] and SuperGLUE [27] set evaluation standards, with limited low-resource language representation.

Existing literature offers (i) robust Portuguese models with absolute performance focus, (ii) strong QA baselines on SQuAD-BR, and (iii) mature PEFT techniques mainly evaluated in English. What is missing is a study combining these, systematically evaluating PEFT and quantization in Brazilian Portuguese QA—this is the gap addressed here.

IV. Methodology

A. Dataset and Model Configurations

We use the Portuguese-translated SQuAD v1 dataset [21], with 87,599 training and 10,570 evaluation QA pairs. Two BERTimbau variants [5] are evaluated: (1) **BERTimbau-Base**: 12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters; (2) **BERTimbau-Large**: 24 layers, 1024 hidden dimensions, 16 heads, 335M parameters. Both pretrained on the brWaC corpus (2.68B tokens).

B. PEFT Configuration

For all PEFT methods: LoRA rank $r = 16$, scaling factor $\alpha = 32$, target modules (Q,K,V,Output in attention), dropout=0.1. Quantized variants use 4-bit NF4 base weights with double quantization and compute dtype bfloat16. Two learning rates explored: standard BERT (4.25e-5) and PEFT-optimized high LR (2e-4), training 2 and 3 epochs. Common parameters: AdamW optimizer [28], [29], weight decay=0.01, batch size=16 (Base)/8 (Large), max sequence length 384, gradient clipping norm 1.0.

C. Computational Infrastructure

Experiments executed on: (1) **Primary GPU**: NVIDIA RTX A4500 20GB for BERTimbau-Large and full fine-tuning, (2) **Secondary GPU**: NVIDIA RTX 2080 Ti 8GB for BERTimbau-Base and quantized method validation. Software: CUDA 12.2, PyTorch 2.1.0, Transformers 4.36.0, PEFT 0.7.1, bitsandbytes 0.41.0. This setup represents moderate academic environments, showing PEFT viability outside high-performance clusters. Training BERTimbau-Large with QLoRA on RTX 2080 Ti validates democratization via quantization.

V. Experimental Results

A. Performance on BERTimbau-Base

Tables I and II report BERTimbau-Base results for different learning rates and epochs. Full fine-tuning baseline achieves best performance with lr=4.25e-5, 2 epochs (F1=82.79, EM=70.91).

TABLE I
BERTimbau-Base with high learning rate (2e-4)

Method	Epoch	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

The contrast between learning rates is dramatic. With lr=2e-4, LoRA reaches F1=78.01 (94.2% of the baseline), whereas with lr=4.25e-5 it only achieves F1=71.81 (86.7% of the baseline), a difference of **+6.20 F1 points**. This pattern is consistent across all PEFT methods. Quantized variants completely collapse with the standard learning rate (QLoRA: F1=53.52, QDoRA: F1=54.10), falling more than 20 points below the baseline.

Among PEFT techniques with lr=2e-4, LoRA and DoRA tie in performance (F1=78.01) at 2 epochs, but LoRA is 27.8% faster. Quantized variants show significant degradation: QLoRA reaches F1=73.23 (-9.56 points vs. baseline) and QDoRA F1=74.41. This sensitivity suggests that BERTimbau-Base loses critical information during 4-bit compression.

TABLE II
BERTimbau-Base with standard learning rate (4.25e-5)

Method	Epoch	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

B. Performance on BERTimbau-Large

Tables III and IV present complete results for BERTimbau-Large. The baseline achieves its best performance with lr=4.25e-5 at 2 epochs (F1=84.86, EM=73.00). Notably, the baseline with lr=2e-4 completely collapses (F1=3.02), showing that high learning rates are detrimental for full fine-tuning but essential for PEFT.

TABLE III
BERTimbau-Large with high learning rate (2e-4)

Method	Epoch	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

*Estimated

time

TABLE IV
BERTimbau-Large with standard learning rate (4.25e-5)

Method	Epoch	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

*Estimated

time

The impact of learning rate on Large is consistent with Base but more pronounced. With lr=2e-4, LoRA reaches F1=81.32 (95.8% of baseline), whereas with lr=4.25e-5 it only achieves F1=75.65 (89.1%), a difference of **+5.67 F1 points**. Quantized variants also exhibit severe degradation with the standard lr: QLoRA drops from F1=80.03 to 68.23 (-11.80 points).

BERTimbau-Large demonstrates greater resilience to PEFT techniques than Base. With lr=2e-4 at 2 epochs, LoRA retains 95.8% of baseline performance versus 94.2% in Base, with a 73.5% reduction in training time. Crucially, quantized variants in Large show much less degradation than in Base. QLoRA reaches F1=80.03 (94.3% of baseline, -4.83 points), compared to F1=73.23 (88.5%, -9.56 points) in Base. This 2.0× difference in degradation confirms that larger models are significantly more robust to aggressive quantization.

A remarkable finding is the collapse of the baseline with lr=2e-4 (F1=3.02), in contrast with the success of PEFT methods under the same learning rate. This suggests that the low-rank structure of LoRA/DoRA acts as an implicit regularizer, preventing training divergence.

C. Critical Impact of Learning Rate

The exhaustive results reveal that learning rate is the most decisive factor for PEFT success, outweighing the choice of method, number of epochs, or even the application of quantization. Fig. 1 visualizes this dramatic contrast.

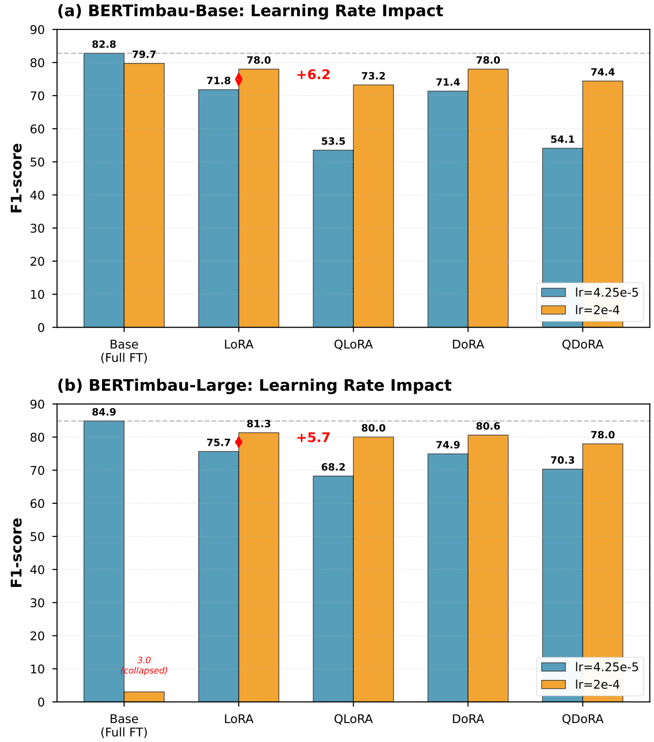

Fig. 1. Impact of learning rate on F1 performance at 2 epochs. High learning rates (2e-4) are essential for PEFT, while collapsing full fine-tuning in Large.

In BERTimbau-Base with LoRA, lr=2e-4 reaches F1=78.01 versus F1=71.81 with lr=4.25e-5, an improvement of +6.20 points (+8.6% relative). In Large, the improvement is +5.67 points (+7.5% relative). For quantized methods, the impact is more dramatic: QLoRA improves +19.71 points in Base (53.52 → 73.23) and +11.80 in Large (68.23 → 80.03). With standard learning rate, quantized methods collapse, falling more than 20–30 points below the baseline.

D. Model Size Scalability

Results show that PEFT techniques scale favorably with model size. BERTimbau-Large exhibits: (1) lower performance degradation with LoRA (95.8% vs. 94.2% in Base), (2) higher resilience to quantization (loss of 4.83 vs. 9.56 points), and (3) greater computational efficiency. This trend suggests that redundancy in larger models allows low-rank adaptations to capture necessary transformations more effectively.

E. Learning Rate for PEFT: Critical Finding

Learning rate is the most critical factor for PEFT. With high lr (2e-4), LoRA achieves competitive performance, but with standard lr (4.25e-5), performance collapses. The +6.20 point difference in Base and +5.67 in Large represent +8.6% and +7.5% relative improvements. For quantized methods, QLoRA gains +19.71 points in Base with high lr.

While lr=2e-4 is essential for PEFT, it is catastrophic for full fine-tuning in Large (F1=3.02 vs. 84.86). This opposite behavior reflects fundamentally different optimization dynamics. In full fine-tuning, every parameter contributes directly in a 335M-dimensional space. In contrast, LoRA operates in a compressed subspace of 13.4M dimensions, where updates propagate through the factorization $\Delta W = BA$. This compression requires more aggressive gradients to compensate for reduced expressive capacity. The low-rank structure also acts as an implicit regularizer: even with high lr, updates are constrained to the rank-r subspace, preventing divergence.

F. LoRA vs. DoRA Comparison

DoRA achieves similar or slightly lower performance than LoRA in both models but with significant time overhead (+27.8% in Base, +28.6% in Large). The explicit magnitude-direction decomposition, though theoretically more expressive, does not yield practical improvements for QA in Portuguese. For practical applications, LoRA is preferable due to simplicity and equal or superior efficiency.

G. Impact of Quantization

4-bit quantization introduces model-size-dependent degradation. In Base, QLoRA loses 11.5% of baseline performance (9.56 F1 points), while in Large it loses only 5.7% (4.83 points). This 2.0× difference suggests smaller models have less redundancy to absorb quantization noise. For memory-constrained applications, QLoRA with Large offers a better trade-off, achieving competitive performance (F1=80.03, 94.3% of baseline) with minimal memory requirements.

H. Practical Recommendations

For 8GB VRAM: QLoRA + Large (F1=80.03, 94.3% baseline), validated on RTX 2080 Ti, democratizes access to large models.

For 16–20GB VRAM: LoRA + Large (F1=81.32, 95.8% baseline, 73.5% training time reduction), common in academic labs.

For small models: Avoid quantization (loss 9.56 vs 4.83). LoRA without quantization is preferable.

For deployment: LR is critical: 2e-4 (not 4.25e-5). DoRA is not recommended (+28% overhead).

Hardware feasibility: This study demonstrates that advanced PEFT techniques are feasible on standard infrastructure. BERTimbau-Large was successfully fine-tuned with QLoRA on 8GB GPUs, without specialized hardware. This accessibility is crucial for democratizing NLP research for Brazilian Portuguese under limited-resource contexts.

VI. Conclusion

This work presents the first systematic evaluation of PEFT and quantization techniques applied to BERTimbau in QA, executed on accessible hardware. Key findings reveal that: (1) LoRA reaches 95.8% of full fine-tuning performance with 73.5% training time reduction, (2) quantized techniques maintain competitive accuracy (QLoRA: 94.3% of baseline), representing the first quantization application to BERTimbau, (3) significantly higher learning rates (2e-4) are essential for PEFT, improving +6.20 F1 points, (4) larger models exhibit greater resilience to quantization (loss of 4.83 vs. 9.56 points), and (5) DoRA offers no practical advantage over LoRA for QA in Portuguese.

Contributions include: (i) first quantization study applied to BERTimbau, (ii) evidence-based practical recommendations for PEFT selection, (iii) identification of learning rate as a frequently underestimated critical factor, and (iv) validation of feasibility on accessible hardware (8–20GB).

Future work includes: (1) exploration of more aggressive quantization (3-bit, 2-bit) with Large, (2) evaluation of PEFT on other NLP tasks in Portuguese, (3) analysis of different LoRA configurations (rank, α), and (4) extension to Brazilian autoregressive models such as Sabiá and Tucano.

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] 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

[22] 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

[23] 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/

[24] 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/

[25] P. V. Quinta de Castro, N. Félix Felipe da Silva, and A. da Silva Soares, “Portuguese named entity recognition using lstm-crf,” in *Computational*

Processing of the Portuguese Language, A. Villavicencio, V. Moreira, A. Abad, H. Caseli, P. Gamallo, C. Ramisch, H. Gonçalo Oliveira, and G. H. Paetzold, Eds. Cham: Springer International Publishing, 2018, pp. 83–92.

[26] 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

[27] 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

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

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