Modestos e Sustentáveis: O Ajuste Eficiente Beneficia Modelos de Língua de Menor Escala em Português?

Gabriel Assis, Arthur Vasconcelos, Lívia de Azevedo, Mariza Ferro, Aline Paes

Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brasil

{assisgabriel,athurbittencourt,liviaazevedosilva}@id.uff.br,
{mariza,alinepaes}@ic.uff.br

Abstract. *Language Models have established new performance standards in text-based tasks. Yet, these models require substantial amounts of data and computational power. This article investigates Parameter Efficient Fine-Tuning (PEFT) techniques, specifically LoRA and GreenTrainer, on Portuguese specialized models OPT-PTBR and PTT5. We aim to evaluate whether PEFT maintains model performance while reducing the financial and environmental costs associated with intensive resource consumption, even in small-scale models. Our results show that GreenTrainer, in particular, delivers performance comparable to full Fine-Tuning while significantly reducing computational demands.*

Resumo. *Modelos de Língua têm estabelecido novos padrões de desempenho em tarefas textuais. Porém, tais modelos exigem grandes volumes de dados e recursos computacionalis intensivos. Este estudo explora o uso de técnicas de Ajuste Fino Eficiente de Parâmetros (PEFT), especificamente LoRA e GreenTrainer, aplicadas a modelos especializados para o português, OPT-PTBR e PTT5. Almeja-se avaliar se as técnicas de PEFT mantêm o desempenho dos modelos enquanto mitigam os impactos financeiros e ambientais do uso intensivo de recursos, mesmo em modelos menores. Os resultados mostram que o GreenTrainer, particularmente, oferece desempenho competitivo em relação ao Ajuste Fino completo, enquanto reduz significativamente demandas computacionais.*

1. Introdução

Modelos de Língua (MLs) Computacionais têm como objetivo representar componentes da língua humana de forma simplificada usando representações numéricas, mas tentando preservar seus fundamentos léxicos, sintáticos e semânticos [Paes et al. 2024]. No contexto atual do Processamento de Linguagem Natural (PLN), MLs Neurais — baseados em redes neurais — que empregam a arquitetura *Transformer* [Vaswani et al. 2017] destacam-se por alcançarem resultados no estado-da-arte em diversas tarefas [Wolf et al. 2020]. Particularmente, MLs de larga escala (*Large Language Models*, LLMs) [Zhao et al. 2023, Paes et al. 2024] estabeleceram novos padrões para tarefas generativas, como a sumarização [Fu et al. 2024]. Tais modelos se caracterizam pelo seu vasto número de parâmetros que possibilitam a observação de habilidades emergentes, ao resolverem tarefas para as quais não foram explicitamente treinados [Paes et al. 2024]. Como consequência, LLMs passaram a ser integrados como componentes de *software* e partes essenciais de agentes de conversação, expandindo seu uso para além dos ambientes acadêmicos e corporativos e tornando-os acessíveis por qualquer indivíduo com um computador.

Dessa forma, aumentou-se a demanda pelo desenvolvimento e acesso de LLMs, acompanhados por um crescimento expressivo no número de parâmetros desses modelos [Maslej et al. 2024]. Contudo, o aumento em larga escala de parâmetros apresenta desafios notáveis, incluindo a necessidade de vastos volumes de dados e um intenso consumo de recursos computacionais [Zhao et al. 2023]. Neste cenário, a Inteligência Artificial Verde (IA Verde) desponta como uma área dedicada a elucidar e reduzir os impactos computacionais — tanto ambientais, como socioeconômicos — do desenvolvimento de soluções em IA [Schwartz et al. 2020]. Atualmente, o desenvolvimento e a pesquisa em modelos de língua são dominados por entidades privadas, e com uma concentração significativa nos Estados Unidos, União Europeia e China [Maslej et al. 2024]. Essa concentração representa um entrave, pois limita a diversificação de pesquisa em outras regiões, como o Brasil, que enfrentam restrições de recursos. Além disso, a sustentabilidade ambiental emerge como uma questão crítica, dado, por exemplo, o alto uso de tempo em GPUs para treinamento e operação de MLs, que tem como consequências um elevado consumo energético e seu equivalente em emissões de dióxido de carbono (CO_2eq) e uso de água potável [Li et al. 2023].

No contexto de adaptação de MLs, técnicas como o Ajuste Fino (*Fine-Tuning*) e, mais ainda, o Ajuste Fino Eficiente de Parâmetros (*Parameter Efficient Fine-Tuning*, PEFT) [Xu et al. 2023] emergem como abordagens para adaptar LLMs de forma a aliviar essas limitações. Ambas as abordagens aproveitam o conhecimento previamente codificado em MLs Pré-Treinados (*Pre-trained Language Models*, PLMs) [Ding et al. 2023] e os adaptam para domínios ou tarefas específicas. Entretanto, enquanto a primeira abordagem pode alterar todos os parâmetros do modelo pré-treinado, a segunda abordagem foca na adaptação considerando explicitamente a limitação de recursos. Todavia, diversos métodos de PEFT dependem da seleção de parâmetros a serem alterados, o que pode acarretar em degradação de desempenho [Yang et al. 2024].

Os métodos de PEFT são tipicamente avaliados em LLMs com bilhões de parâmetros, sob a premissa de que esses modelos são superparametrizados [Ding et al. 2023]. Embora haja uma motivação natural para reduzir o consumo de recursos por parte desses modelos, sua aplicação em grande escala, mesmo que de forma mais eficiente, não elimina completamente as barreiras impostas ao uso de MLs dessa magnitude. Surge, então, uma questão relevante: *quais seriam os impactos da aplicação de técnicas de PEFT em modelos de menor escala em relação a sua capacidade de realizar tarefas específicas?* Adicionalmente, o português destaca-se como uma língua diversificada, apresentando particularidades estruturais significativas, como a relação de ordem das palavras e as variações nas desinências, que podem alterar o significado de uma frase [Kato et al. 2023]. Nesse contexto, outra questão importante se apresenta: *a aplicação de técnicas de PEFT em modelos de menor escala para o português afetaria negativamente o desempenho e a representação do idioma?*

Para responder tais questões, este artigo contribui com uma avaliação entre a abordagem de ajuste fino completo e técnicas de PEFT, especificamente *Low-Rank Adaptation* (LoRA) [Hu et al. 2022] e GreenTrainer [Huang et al. 2024], em dois PLMs específicos para o português: OPT-PTBR[1], com 125 milhões de parâmetros, e PTT5-base [Carmo et al. 2020], com 223 milhões de parâmetros. Nossos resultados demons-

[1] https://huggingface.co/monilouise/opt125M_portuguese

tram que as técnicas eficientes produzem desempenhos competitivos em relação ao ajuste fino completo, mesmo em modelos de menor escala. Notavelmente, a técnica GreenTrainer apresentou resultados com menor degradação e, em alguns casos, até superiores ao ajuste fino completo. Com essa análise, buscamos contribuir para a atenuação dos impactos socioeconômicos e ambientais do treinamento de MLs, sem deixar de considerar as particularidades do idioma português.

2. Fundamentação Teórica

Esta seção visa elucidar conceitos fundamentais tratados no trabalho e essenciais no contexto de ajuste de MLs, especificamente acerca de PLMs e métodos de PEFT.

2.1. Ajuste de Modelos de Língua Pré-treinados

Os PLMs são modelos que passam por uma etapa chamada de pré-treinamento, cujo objetivo é incorporar informações linguísticas relevantes a partir de um grande volume de *corpora*. Todavia, esses modelos podem não representar adequadamente informações específicas de certos domínios ou tarefas não abordadas durante o pré-treinamento. Para tratar dessa questão, adota-se amplamente o ajuste fino dos PLMs, no qual os pesos dos modelos são atualizados para tarefas ou domínios particulares por meio do treinamento sobre um novo conjunto de dados específico, tipicamente na tarefa final pretendida. Dessa forma, é possível aproveitar o conhecimento previamente codificado sem a necessidade de repetir a etapa de pré-treinamento, realizando um processo direcionado e geralmente menos oneroso [Paes et al. 2024].

2.2. Ajuste Fino Eficiente de Parâmetros

O conjunto de técnicas de PEFT reduz a demanda por recursos computacionais para ajuste de PLMs. Esses métodos são divididos por [Xu et al. 2023] em aditivo, parcial, reparametrizado, unificado e híbrido. O ajuste aditivo introduz uma quantidade menor de parâmetros adicionais ajustáveis, evitando o ajuste dos parâmetros próprios do modelo pré-treinado. O ajuste parcial atualiza apenas um subconjunto dos parâmetros pré-treinados. A reparametrização utiliza transformações de baixo posto da Álgebra Linear para reduzir o número de parâmetros treináveis. O método unificado propõe um *framework* coeso que simplifica a integração de técnicas de ajuste fino, garantindo consistência e eficiência na adaptação dos modelos. Por fim, o método híbrido combina diversas técnicas de PEFT. Em comum, todos os métodos ajustam um número reduzido de parâmetros dos MLs.

Dentre todas essas técnicas, o método do tipo reparametrizado LoRA [Hu et al. 2022] se destaca como um dos métodos de PEFT mais utilizados para o ajuste de modelos em diferentes tarefas ao proporcionar consistentemente a redução no número de parâmetros treináveis e consequente redução na demanda de memória [Zhao et al. 2024a, Yang et al. 2024]. Essa estratégia utiliza matrizes adicionais de baixo posto $\mathbf{A}$ e $\mathbf{B}$, que substituem a matriz de pesos original $\mathbf{W}$. A computação final dos modelos é realizada por meio da expressão $\mathbf{W} + \mathbf{A} \times \mathbf{B}$, permitindo a adaptação dos pesos com uma quantidade significativamente menor de recursos computacionais.

Embora eficaz, a LoRA ainda requer a computação dos gradientes de ativação durante a etapa de *backpropagation* no treinamento de modelos, o que limita seu potencial

máximo de redução de recursos. A estratégia GreenTrainer [Huang et al. 2024] surge como uma alternativa que visa reduzir diretamente as operações necessárias para ajustes dos modelos, sem desconsiderar a *backpropagation*. Ela seleciona tensores específicos para ajuste a cada época de treinamento, com base na importância de cada tensor para a diminuição da *loss*, caracterizando-se assim como uma técnica de ajuste parcial. Além disso, ela permite a configuração do hiperparâmetro ρ, que determina a porcentagem de operações mantidas em relação ao ajuste fino completo.

Desse modo, ao considerar uma técnica consolidada e amplamente reconhecida como a LoRA, e uma abordagem emergente e competitiva, como o GreenTrainer, este estudo visa realizar uma avaliação inicial acerca do impacto dessas abordagens no desempenho de PLMs de menor escala em tarefas finais, bem como na redução de seus custos e impactos computacionais.

3. Trabalhos Relacionados

Trabalhos recentes têm desenvolvido MLs específicos para o português utilizando métodos eficientes. Como ilustração, [Carmo et al. 2020] realizaram tanto o ajuste completo de parâmetros quanto o ajuste restrito aos *embeddings* do vocabulário — um método parcial — no treinamento de um ML voltado para o português. Os resultados indicam que, embora competitivo, o ajuste restrito aos *embeddings* é inferior ao ajuste completo. Além disso, os estudos de [Garcia et al. 2024] e [Cabral et al. 2024] introduziram LLMs ajustados especificamente para tarefas em português baseados na arquitetura Llama [Touvron et al. 2023], empregando a técnica de reparametrização LoRA.

Outros trabalhos avaliam o impacto de técnicas de PEFT sobre o desempenho de PLMs. [Yang et al. 2024] comparam o ajuste fino completo a técnicas como LoRA, Prefix-tuning [Li and Liang 2021] e o uso de adaptadores [Houlsby et al. 2019] em modelos de menor escala da arquitetura BERT [Devlin et al. 2019] em tarefas não generativas, destacando o desempenho da estratégia LoRA e a competitividade das demais estratégias de PEFT em relação ao ajuste completo nesse contexto. Contudo, tratando-se da avaliação realizada sobre modelos generativos da arquitetura Llama, as técnicas baseadas em LoRA se sobressaem. Resultados similares são reportados em [Huang et al. 2024], que, ao proporem a estratégia GreenTrainer, realizaram uma avaliação comparativa com outras técnicas de PEFT, concluindo por sua competitividade. O estudo avaliou PLMs multilíngues ou predominantemente voltados ao inglês, com parâmetros variando entre 350 milhões e 7 bilhões, revelando o potencial da aplicação de técnicas eficientes até mesmo nos modelos com menor número de parâmetros.

No melhor de nosso conhecimento, não há, ainda, trabalho que avalie o uso da abordagem GreenTrainer para MLs em português. Adicionalmente, nenhum dos trabalhos mencionados apresenta uma análise comparativa que considere a relação do desempenho de MLs de menor escala no idioma e o impacto de seu consumo em termos de tempo e CO_2eq. Desse modo, este estudo visa oferecer novas perspectivas que abordem tanto a eficácia preditiva de modelos, quanto os custos associados a sua etapa de ajuste.

4. Avaliação de Técnicas de Ajuste Eficiente

Esta seção detalha os MLs avaliados, a tarefa de PLN selecionada e as métricas de avaliação adotadas para investigar o impacto das técnicas LoRA e GreenTrainer tanto

no desempenho textual quanto no consumo computacional. Destaca-se que o ajuste fino completo dos parâmetros dos modelos foi adotado como *baseline*.

4.1. Modelos de Língua Selecionados

A seleção de MLs para tarefas específicas é influenciada pelo número de parâmetros e pelo *corpus* de pré-treinamento, fatores cruciais para a viabilidade de execução em diferentes plataformas de *hardware* e para as capacidades de geração de texto do modelo. Modelos maiores geralmente demandam mais recursos computacionais, enquanto o *corpus* de pré-treinamento determina a adequação do modelo às necessidades da tarefa [Freitas 2024]. No contexto de recursos limitados, foram escolhidos dois modelos: o **OPT-PTBR**[2], com 125 milhões de parâmetros, baseado na arquitetura Open Pretrained Transformer (OPT) [Zhang et al. 2022] e adaptado para o português do Brasil, e o **PTT5-base** [Carmo et al. 2020], com 223 milhões de parâmetros, utilizando a arquitetura T5 [Raffel et al. 2020] e pré-treinado com um *corpus* de páginas *web* em português do Brasil. A escolha de modelos menores alinha-se com a necessidade de operar em ambientes com recursos limitados, mantendo a avaliação da qualidade da geração de textos em português.

4.2. A Tarefa de PLN Aplicada: Sumarização

Para garantir a compatibilidade com a implementação pública do GreenTrainer[3], a tarefa de sumarização textual foi selecionada. A sumarização por meio de MLs consiste em condensar as informações de um texto, gerando uma nova versão que preserva de forma concisa o conteúdo essencial do original. Essa tarefa é amplamente estudada em PLN, incluindo no contexto do português brasileiro [Paiola 2022, Pontes et al. 2022, Feltrin et al. 2023], com LLMs recentemente estabelecendo novos padrões de geração [Souza et al. 2024]. Fatores como a coocorrência de termos relevantes e a fidelidade entre texto original e gerado são importantes para determinar a qualidade de um resumo. Igualmente relevantes são aspectos como a aderência a formalidade e precisão gramatical pretendidos. Por exemplo, no contexto jornalístico, resumos de notícias políticas podem exigir um nível de formalidade distinto daquele necessário para resumos de eventos recentes em um *reality show* popular, embora, em ambos os casos, a correção gramatical seja tipicamente fundamental. Assim, a tarefa de sumarização posiciona-se apropriadamente para a avaliação da aplicação de técnicas de ajuste de modelos, uma vez que a adequação a contextos e domínios específicos é fundamental para garantir a qualidade das gerações textuais [Paes et al. 2024].

4.3. Métricas de Avaliação

Com o objetivo de avaliar de forma integrada a qualidade do desempenho generativo e os custos e impactos computacionais, três grupos de métricas foram usados na análise de geração de sumários. O primeiro grupo visa medir a aderência dos resumos gerados em relação aos textos de referência e é composto pelas métricas ROUGE [Lin 2004] e BERTScore [Zhang et al. 2020]. A métrica ROUGE, amplamente utilizada nesse contexto, compara a sobreposição de n-gramas entre o sumário automático e a referência, enquanto o BERTScore utiliza modelos de língua baseados em BERT para avaliar a similaridade semântica entre os textos. Neste estudo, a métrica ROUGE é apresentada pela média

[2] https://huggingface.co/monilouise/opt125M_portuguese
[3] https://github.com/pittisl/GreenTrainer/

de suas variantes, ROUGE-1, ROUGE-2, ROUGE-L e ROUGE-S [Souza et al. 2024], que se diferenciam na forma de computar os n-gramas, sendo os resultados expressos em valores percentuais. O BERTScore, por sua vez, é expresso em termos da sua componente F1. O segundo grupo de métricas visa mensurar explicitamente o impacto e o consumo de recursos associados ao ajuste dos modelos, incluindo a contagem do número de (peta) operações de ponto flutuante por segundo (PFLOPS), que quantifica as operações aritméticas necessárias para o ajuste, o tempo de treinamento dos modelos e a quantidade equivalente de CO_2 emitida durante o processo, estimada pela ferramenta disponível por [Lacoste et al. 2019]. Por fim, o terceiro grupo aproveita do extenso conjunto de métricas fornecidas pelo portal NILC-*metrix* [Leal et al. 2023][4] para avaliar a qualidade de escrita dos textos gerados. Essas métricas extraem valores de diversos indicadores linguísticos para avaliar informações sobre morfossintaxe, coesão e coerência.

5. Experimentos

Esta seção apresenta os experimentos conduzidos, detalhando as configurações utilizadas e os resultados obtidos.

5.1. Configurações Experimentais

Hiperparâmetros Considerando a premissa de recursos limitados, os modelos foram treinados por apenas uma época, com uma taxa de aprendizado de $2 \cdot 10^{-5}$ e um tamanho de lote de 4. Para a tarefa de sumarização, foram definidos: max_input_length de 512, max_output_length de 128, $repetition\_penalty$ de 2,5 e $length\_penalty$ de 1,0. No que se refere aos parâmetros do LoRA, utilizou-se $r = 8$, $lora\_alpha = 32$ e uma taxa de $dropout$ de 0,1. O GreenTrainer foi testado com ρ de 0,5 e 0,7, e implementado conforme [Huang et al. 2024]. Também ao encontro desse trabalho, o modelo OPT-PTBR foi configurado com a estrutura "TL;DR" para sumarização, enquanto o modelo PTT5 usou o prefixo "sumarize: [sequência de entrada]". Por fim, o BERTScore foi computado utilizando o modelo BERT multilingual[5], dada a incompatibilidade da métrica com um modelo próprio para o português.

Conjunto de Dados A tarefa de sumarização ocorreu com a base RecognaSumm [Paiola et al. 2024]. Esse conjunto possui origem diversificada, sendo composto por notícias de diferentes fontes de informação. Tal diversidade resulta em uma coleção de documentos que abrangem uma variedade de tópicos e estilos jornalísticos. Ademais, o RecognaSumm contém cerca de 135 mil instâncias em que, para os propósitos deste trabalho, foram selecionadas apenas as colunas referentes ao texto da notícia e ao sumário, esse último servindo como referência padrão nas métricas de avaliação. Adota-se a subdivisão pré-estabelecida do conjunto de dados, de 81,2 mil instâncias para treinamento e 27,1 mil para validação e teste cada.

5.2. Resultados Experimentais

A Tabela 1 combina os resultados do primeiro e do segundo grupo de métricas avaliados. A porcentagem indica a variação positiva ou negativa em relação ao ajuste fino completo, com resultados com diferença percentual inferior a 1% marcados com 0%. Por fins de simplificação, as configurações de ρ para o GreenTrainer são denotadas GT-ρ.

[4] http://fw.nilc.icmc.usp.br:23380/metrixdoc

[5] https://huggingface.co/google-bert/bert-base-multilingual-cased

Tabela 1. Comparação de eficiência, impacto ambiental e métricas textuais.

Modelo	Estratégia	PFLOPS	Tempo (h)	CO_2eq (kg)	ROUGE	BERTScore
OPT-PTBR *(125M params)*	Ajuste fino	16,59	3,00	0,24	7,48	0,652
	GT-0.5	8,18 (51%↓)	1,45 (52%↓)	0,12 (50↓%)	4,60 (39%↓)	0,662 (2%↑)
	GT-0.7	11,61 (30%↓)	2,17 (28%↓)	0,17 (29↓%)	7,94 (6%↑)	0,682 (5%↑)
	LoRA	11,06 (33%↓)	2,28 (24%↓)	0,18 (25↓%)	7,23 (3%↓)	0,672 (3%↑)
PTT5 *(220M params)*	Ajuste fino	15,67	3,73	0,30	27,82	0,742
	GT-0.5	8,76 (44%↓)	2,38 (36%↓)	0,19 (37↓%)	27,16 (2%↓)	0,739 (0%↓)
	GT-0.7	11,50 (27%↓)	2,87 (23%↓)	0,23 (23↓%)	27,56 (1%↓)	0,741 (0%↓)
	LoRA	10,45 (33%↓)	2,61 (30%↓)	0,21 (30↓%)	26,20 (6%↓)	0,734 (1%↓)

Os resultados indicam que a estratégia GT-0.7 apresentou ou a menor degradação, ou uma melhora no desempenho textual em comparação com o ajuste fino em todos os casos avaliados. Em termos de desempenho computacional, seus resultados são próximos aos da LoRA, embora ligeiramente inferiores. Observa-se que a configuração GT-0.5, a mais eficiente em termos de consumo, apresentou uma queda significativa nos resultados generativos para o OPT-PTBR na métrica ROUGE. No entanto, essa mesma configuração não resultou em grandes quedas para o modelo PTT5, indicando que a robustez inerente do modelo deve ser considerada ao aplicar estratégias de eficiência drástica. Na verdade, essa configuração foi superior à estratégia LoRA para esse modelo.

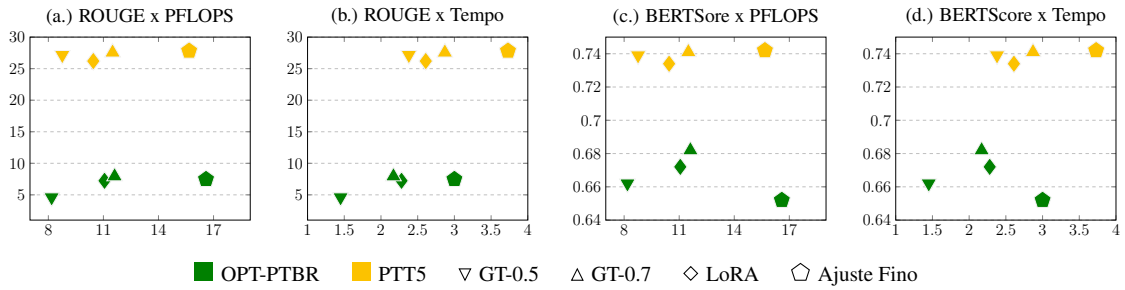

Figura 1. Comparativo entre os desempenhos computacionais e textuais.

A Figura 1 contrasta as métricas textuais com as medidas de desempenho. Essa comparação reforça a estratégia GT-0.7 como a que gera resultados mais próximos do ajuste fino, seguida pela estratégia LoRA. Fica evidente também a leve superioridade da economia da estratégia LoRA em relação à GT-0.7. Em termos de economia computacional, a estratégia LoRA se posiciona consistentemente entre as configurações de GT, embora, em termos de resultados generativos, seja inferior à GT-0.5 para o modelo PTT5. Além disso, particularmente para o modelo OPT-PTBR, os resultados de BERTScore obtidos pelo ajuste eficiente foram superiores ao ajuste completo. Por fim, a Figura 1 também demonstra a superioridade geral do modelo PTT5 na execução da tarefa, ressaltando o impacto que a escolha adequada do PLM pode implicar.

A Tabela 2 apresenta a distância euclidiana média, calculada com base em cinco conjuntos de métricas do NILC-*metrix*, entre uma amostra de 100 sumários gerados para cada configuração avaliada e suas respectivas referências. Antes do cálculo, os valores foram normalizados para o intervalo de 0 a 1. Os melhores resultados estão destacados em negrito, enquanto os segundos melhores estão sublinhados. De modo geral, os valores semelhantes observados dentro do mesmo modelo, independentemente da estratégia

Tabela 2. Métricas em sintaxe, morfologia e semântica do conjunto NILC-*metrix*.

Modelo	Estratégia	Coesão Referencial	Coesão Semântica	Informações Semânticas	Complexidade Sintática	Informações Morfossintáticas
OPT-PTBR *(125M params)*	Ajuste fino	0,259	0,667	0,587	0,276	1,150
	GT-0.5	0,269	0,679	0,595	0,255	1,143
	GT-0.7	0,252	0,652	0,560	0,281	1,058
	LoRA	0,279	0,760	0,600	0,264	1,158
PTT5 *(220M params)*	Ajuste fino	0,340	0,491	0,513	0,253	**0,838**
	GT-0.5	0,310	0,505	**0,481**	**0,237**	0,853
	GT-0.7	0,326	**0,446**	0,492	0,267	0,851
	LoRA	**0,247**	0,692	0,560	0,258	0,962

de ajuste, indicam que as técnicas de PEFT não comprometem significativamente a capacidade de escrita dos modelos de língua quando comparadas ao ajuste fino. Notavelmente, as configurações do GT obtiveram os melhores resultados em várias ocorrências. No entanto, uma avaliação comparativa entre os modelos revela que o PTT5 consistentemente apresenta desempenho superior, especialmente na avaliação de Informações Morfossintáticas. Esse resultado pode estar relacionado à etapa de pré-treinamento, mais robusta nesse modelo, sugerindo que uma execução adequada dessa fase possibilita uma estratégia de ajuste mais eficiente. No entanto, uma dualidade surge, pois uma etapa de pré-treinamento mais robusta pode resultar em custos mais elevados. De maneira geral, esses resultados corroboram os anteriormente descritos, indicando que, além da escolha da estratégia de ajuste, a seleção do modelo mostra-se crucial.

6. Considerações Finais

Este trabalho conduziu experimentos com estratégias de ajuste fino eficiente, empregando dois modelos de menor escala treinados em português para a tarefa de sumarização textual. Os resultados indicam que a estratégia do GreenTrainer é competitiva em relação à estratégia já estabelecida LoRA. Dependendo da escolha do parâmetro ρ, a estratégia pode, inclusive, alcançar um equilíbrio superior entre a degradação de desempenho e o ganho de eficiência computacional. Além disso, os resultados revelam que a aplicação de estratégias eficientes pode implicar degradações significativas, dependendo da escolha do modelo. Trabalhos futuros incluem avaliar essas estratégias em outros modelos e tarefas, visando obter melhores indicativos sobre a generalização, além de considerar novas estratégias como LoRETTA [Yang et al. 2024] e GaLore [Zhao et al. 2024b].

Por fim, visando à transparência, explicitamos os custos totais desta pesquisa, totalizando R$1.642,48 em uso de recursos em nuvem. Os experimentos, realizados na *Google Cloud Platform* na região *us-central1*, resultaram em emissões estimadas de 14,52 kgCO_2eq, com 364 horas de computação em duas GPUs T4 (TDP de 70W) e uma eficiência de carbono de 0,57 kgCO_2eq/kWh.

Agradecimentos

Os autores agradecem ao financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), bolsa 307088/2023-5, da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), processos SEI-260003/002930/2024, SEI-260003/000614/2023, e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) — Código Financeiro 001. Também agradecem aos recursos do programa *Google Cloud Research Credits*, código GCP19980904.

Referências

Cabral, B., Claro, D., and Souza, M. (2024). Exploring Open Information Extraction for Portuguese Using Large Language Models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 127–136.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. *arXiv*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J., and Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Feltrin, G., Vianna, D., and da Silva, A. (2023). Um Estudo sobre Métricas de Avaliação para Sumarização de Acórdãos. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 295–305, Porto Alegre, RS, Brasil. SBC.

Freitas, C. (2024). Dataset e corpus. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 13. BPLN, 2 edition.

Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. (2024). GPTScore: Evaluate as You Desire. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Júnior, A. C., Jodas, D. S., Afonso, L., Guilherme, I. R., Penteado, B. E., and Papa, J. P. (2024). Introducing Bode: A Fine-Tuned Large Language Model for Portuguese Prompt-Based Task. *arXiv preprint arXiv:2401.02909*.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, K., Yin, H., Huang, H., and Gao, W. (2024). Towards Green AI in Fine-tuning Large Language Models via Adaptive Backpropagation. In *The Twelfth International Conference on Learning Representations*.

Kato, M. A., Martins, A. M., and Nunes, J. (2023). *The Syntax of Portuguese*. Cambridge Syntax Guides. Cambridge University Press.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700*.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2023). NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, pages 1–38.

Li, P., Yang, J., Islam, M. A., and Ren, S. (2023). Making AI less "Thirsty': Uncovering and Addressing the Secret Water Footprint of AI models.

Li, X. L. and Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. (2024). Artificial Intelligence Index Report 2024.

Paes, A., Vianna, D., and Rodrigues, J. (2024). Modelos de linguagem. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 15. BPLN, 2 edition.

Paiola, P. H. (2022). Sumarização abstrativa de textos em português utilizando aprendizado de máquina. Mestrado em ciências da computação, Universidade Estadual Paulista Júlio de Mesquita Filho, [s.l.]. Programa de Pós-Graduação em Ciência da Computação.

Paiola, P. H., Garcia, G. L., Jodas, D. S., Correia, J. V. M., Sugi, L. A., and Papa, J. P. (2024). RecognaSumm: A Novel Brazilian Summarization Dataset. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 575–579, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Pontes, L., Oliveira, H., and Boldt, F. (2022). Avaliação de Modelos Neurais para Sumarização de Código-fonte. In *Anais do XLIX Seminário Integrado de Software e Hardware*, pages 140–151, Porto Alegre, RS, Brasil. SBC.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12):54–63.

Souza, J. W. d. C., Cardoso, P. C. F., and Paixão, C. A. (2024). Sumarização automática. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 22. BPLN, 2 edition.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xu, L., Xie, H., Qin, S. J., Tao, X., and Wang, F. L. (2023). Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A critical review and assessment. *CoRR*, abs/2312.12148.

Yang, Y., Zhou, J., Wong, N., and Zhang, Z. (2024). LoRETTA: Low-Rank Economic Tensor-Train Adaptation for Ultra-Low-Parameter Fine-Tuning of Large Language Models. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3161–3176, Mexico City, Mexico. Association for Computational Linguistics.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models. *arXiv*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., Sherstinsky, A., Molino, P., Addair, T., and Rishi, D. (2024a). LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report. *arXiv preprint arXiv:2405.00732*.

Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. (2024b). GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.