CAPACIDADE EXPRESSIVA E LIMITAÇÕES DA LOW-RANK ADAPTATION (LORA) EM MODELOS DE LINGUAGEM DE GRANDE PORTE

Expressive Capacity and Limitations of Low-Rank Adaptation (LoRA) in Large-Scale Language Models

João Roberto Peres Lanza*
Adriane Cavichiolli**

Resumo

Este artigo analisa a eficácia da Low-Rank Adaptation (LoRA) e de suas principais variantes, QLoRA, AdaLoRA e DoRA, na adaptação de Modelos de Linguagem de Grande Porte (LLMs). O estudo foi conduzido por meio de revisão bibliográfica narrativa fundamentada na hipótese da dimensão intrínseca. Foram comparadas as estratégias de otimização, eficiência de memória e capacidade expressiva de cada método. Os resultados indicam que a QLoRA se destaca pela redução do consumo de memória através da quantização em 4 bits, a AdaLoRA melhora a utilização do orçamento de parâmetros por meio da alocação dinâmica de rank, e a DoRA apresenta desempenho mais próximo ao ajuste fino completo ao desacoplar magnitude e direção dos pesos. Conclui-se que a escolha da técnica depende das restrições de hardware e dos requisitos de desempenho da aplicação.

Palavras-chave: LoRA, QLoRA, DoRA, AdaLoRA, LLMs.

Abstract

This article analyzes the effectiveness of Low-Rank Adaptation (LoRA) and its main variants, QLoRA, AdaLoRA, and DoRA, in adapting Large-Scale Language Models (LLMs). The study was conducted through a narrative literature review based on the intrinsic dimension hypothesis. The optimization strategies, memory efficiency, and expressive capacity of each method were compared. The results indicate that QLoRA stands out for reducing memory consumption through 4-bit quantization, AdaLoRA improves parameter budget utilization through dynamic rank allocation, and DoRA shows performance closer to full fine-tuning by decoupling the magnitude and direction of the weights. It is concluded that the choice of technique depends on the hardware constraints and performance requirements of the application.

* Aluno do curso de Tecnologia em Análise e Desenvolvimento de Sistemas, da Faculdade de Presidente Prudente. Email: johnrobertopereslanza@gmail.com.
** Professora orientadora Me. Adriane Cavichiolli em Inteligência Artificial da Faculdade de Presidente Prudente. E-mail: adriane.cavichiolli@cps.sp.gov.br.

***Keywords**: LoRA, QLoRA, DoRA, AdaLoRA, LLMs.*

1. INTRODUÇÃO

O desenvolvimento dos Modelos de Linguagem de Grande Porte (Large Language Models, LLMs) transformou o estado da arte no Processamento de Linguagem Natural (Natural Language Processing, NLP), viabilizando avanços expressivos em tarefas como geração de texto, tradução automática, sumarização e raciocínio lógico (HE et al., 2026). A capacidade desses modelos de capturar representações ricas da linguagem humana decorre, em larga medida, de sua escala: os LLMs contemporâneos possuem dezenas ou mesmo centenas de bilhões de parâmetros treináveis, exigindo infraestruturas computacionais de altíssimo custo para seu treinamento inicial.

Diante dessa realidade, a adaptação de LLMs pré-treinados a tarefas específicas de domínio por meio do Ajuste Fino Completo (Full Fine-Tuning, FFT) torna-se impraticável para a maior parte das organizações e grupos de pesquisa. O FFT exige a atualização simultânea de todos os parâmetros do modelo, implicando consumo de memória equivalente ao processo de treinamento original e impossibilitando o uso de equipamentos de hardware convencional (HU et al., 2021). É nesse contexto que as técnicas de Ajuste Fino Eficiente em Parâmetros (Parameter-Efficient Fine-Tuning, PEFT) emergem como alternativa viável, reduzindo drasticamente o número de parâmetros a serem otimizados sem comprometer de forma significativa o desempenho final do modelo.

Entre as abordagens PEFT, a Adaptação de Baixo Rank (Low-Rank Adaptation, LoRA), proposta por Hu et al. (2021), consolidou-se como a técnica de maior adoção na literatura e na indústria. A LoRA opera congelando os pesos originais do modelo e inserindo matrizes de decomposição de baixo rank treináveis nas camadas de atenção, reduzindo substancialmente o número de parâmetros otimizados sem introduzir latência adicional na inferência. Ainda assim, a literatura recente aponta limitações de sua capacidade expressiva em tarefas de alta complexidade, decorrentes da imposição de um rank fixo e uniforme a todas as camadas do modelo (ZENG; LEE, 2023).

Apesar da ampla adoção da LoRA e do surgimento de variantes voltadas à melhoria da eficiência computacional e da capacidade expressiva dos modelos, os estudos disponíveis costumam concentrar-se na avaliação isolada de cada abordagem. Como resultado, pesquisadores e profissionais podem encontrar dificuldades para identificar qual técnica melhor

atende às exigências de diferentes cenários de aplicação. Nesse contexto, torna-se relevante reunir e comparar criticamente as evidências disponíveis sobre LoRA, QLoRA, AdaLoRA e DoRA, fornecendo uma visão integrada de suas características, vantagens, limitações e contextos de uso.

Para superar essas restrições, variantes arquitetônicas foram propostas. O presente artigo tem como objetivo analisar e comparar três das principais: a Adaptação de Baixo Rank Quantizada (Quantized Low-Rank Adaptation, QLoRA), a Adaptação de Baixo Rank Adaptativa (Adaptive Low-Rank Adaptation, AdaLoRA) e a Adaptação de Baixo Rank Decomposta por Peso (Weight-Decomposed Low-Rank Adaptation, DoRA). Por meio de uma revisão crítica da literatura, busca-se estabelecer um panorama das capacidades expressivas, das limitações empíricas e da eficiência de memória de cada método, a fim de orientar a escolha da abordagem mais adequada a diferentes contextos de adaptação de domínio.

2. FUNDAMENTOS TEÓRICOS E ANÁLISE COMPARATIVA DAS VARIANTES

Esta seção apresenta o embasamento teórico e a análise crítica das metodologias avaliadas. Parte-se da hipótese da dimensão intrínseca, que fundamenta matematicamente as abordagens PEFT, para então examinar a LoRA original e suas três principais variantes: QLoRA, AdaLoRA e DoRA. O objetivo é contrastar as inovações arquitetônicas, as eficiências de memória e as limitações empíricas de cada técnica no contexto da otimização de LLMs.

2.1 A hipótese da dimensão intrínseca e os fundamentos do PEFT

O desenvolvimento das técnicas PEFT assenta-se na hipótese da dimensão intrínseca, formalizada por Aghajanyan et al. (2020). Segundo os autores, modelos pré-treinados superparametrizados possuem uma dimensão intrínseca notavelmente baixa, o que indica que as atualizações de pesos necessárias para a aprendizagem de novas tarefas podem ser representadas de forma eficiente em subespaços de dimensão reduzida. Em termos práticos, isso significa que, mesmo em modelos com bilhões de parâmetros, apenas uma fração relativamente pequena do espaço paramétrico é necessária para capturar as especificidades de uma nova tarefa.

Aghajanyan et al. (2020) corroboram essa hipótese ao demonstrar empiricamente que modelos pré-treinados em larga escala apresentam valores de dimensão intrínseca significativamente menores do que o número total de seus parâmetros, e que o próprio processo

de pré-treinamento reduz progressivamente essa dimensão. Nesse contexto, a capacidade expressiva de um modelo corresponde à sua habilidade de representar funções complexas e capturar as nuâncias da tarefa-alvo, sendo diretamente influenciada pela dimensionalidade do subespaço utilizado para as atualizações de pesos.

A partir dessas evidências, a LoRA (HU et al., 2021) opera congelando os pesos originais do modelo pré-treinado e inserindo matrizes de decomposição de baixo rank treináveis em paralelo às camadas de atenção. Formalmente, dada uma matriz de pesos pré-treinada W_0, a LoRA representa a atualização como o produto de duas matrizes de dimensão reduzida, B e A, tal que a atualização resultante é dada pelo produto BA, com rank r significativamente menor do que as dimensões originais da matriz (HU et al., 2021). Essa estratégia reduz substancialmente os parâmetros otimizados e os requisitos de memória, sem introduzir latência adicional na inferência. Embora eficaz, Zeng e Lee (2023) demonstram que a imposição de um rank fixo e uniforme pode restringir a expressividade do modelo em tarefas que demandam a aprendizagem de padrões não lineares de elevada complexidade, motivando o desenvolvimento das variantes apresentadas nas subseções seguintes.

2.2 Análise das principais variantes da LoRA

Para superar a lacuna de desempenho da LoRA original em relação ao FFT, a comunidade de pesquisa desenvolveu variantes algorítmicas que oferecem compromissos distintos entre eficiência de memória, capacidade expressiva e dinâmica de otimização (HE et al., 2026). As três mais representativas na literatura são QLoRA, AdaLoRA e DoRA, cada uma introduzindo inovações arquitetônicas específicas para endereçar limitações distintas da abordagem original.

2.2.1 Adaptação de Baixo Rank Quantizada (QLoRA)

Proposta por Dettmers et al. (2023), a QLoRA foi concebida com o objetivo central de maximizar a eficiência de uso de memória em GPUs. O método combina a adaptação de baixo rank com a quantização dos pesos do modelo base para 4 bits, introduzindo três inovações técnicas fundamentais: o tipo de dado NF4 (NormalFloat 4), otimizado para distribuições normais; a quantização dupla (double quantization), que aplica uma segunda etapa de compressão aos próprios coeficientes de quantização; e o uso de otimizadores paginados, que transferem dinamicamente estados do otimizador para a memória de acesso aleatório (Random

Access Memory, RAM) durante picos de consumo de memória de acesso aleatório de vídeo (Video Random Access Memory, VRAM), evitando erros de alocação durante o treinamento.

Os autores demonstraram que, com esse conjunto de técnicas, torna-se possível o ajuste fino de modelos de até 65 bilhões de parâmetros em uma única GPU com 48 GB de VRAM, mantendo desempenho equivalente ao processamento em precisão de 16 bits (DETTMERS et al., 2023). Em avaliações conduzidas no domínio médico por Toksoz e Isik (2026), a QLoRA demonstrou resultados competitivos mesmo sob severas restrições de hardware, confirmando sua adequação para contextos com recursos computacionais limitados. Essa capacidade posiciona a QLoRA como a principal ferramenta para a democratização do ajuste fino de LLMs em hardware de consumo, sendo especialmente indicada para laboratórios acadêmicos e instituições sem acesso a infraestrutura computacional de larga escala.

2.2.2 Adaptação de Baixo Rank Adaptativa (AdaLoRA)

Enquanto a LoRA padrão atribui um rank fixo de forma uniforme a todas as matrizes de peso, a AdaLoRA, proposta por Zhang et al. (2023), aborda diretamente a ineficiência dessa estratégia. A premissa central é que diferentes camadas e módulos de um LLM contribuem de forma desigual para o desempenho em tarefas específicas; portanto, a alocação uniforme de parâmetros treináveis representa um desperdício do orçamento de parâmetros disponível.

Para contornar esse problema, a AdaLoRA parametriza as atualizações por meio de decomposição em valores singulares (Singular Value Decomposition, SVD) e introduz um mecanismo de alocação dinâmica de rank. Durante o treinamento, são calculadas pontuações de importância para cada valor singular com base em sua magnitude e na sensibilidade da função de perda em relação a cada componente. Os valores singulares menos relevantes são progressivamente descartados, enquanto maior capacidade representacional é redirecionada para as camadas que mais contribuem para o desempenho na tarefa-alvo (ZHANG et al., 2023).

Conforme analisado por Toksoz e Isik (2026) em comparação aplicada ao domínio médico, a AdaLoRA demonstra ganhos consistentes na eficiência alocativa de parâmetros em relação à LoRA padrão, especialmente em tarefas onde o orçamento de parâmetros treináveis é altamente restrito. Uma limitação empírica relevante, contudo, é que o mecanismo adaptativo demanda ciclos de treinamento mais extensos para que as pontuações de importância se estabilizem, o que pode tornar a abordagem menos vantajosa em contextos nos quais o tempo de treinamento é uma variável crítica (TOKSOZ; ISIK, 2026).

2.2.3 Adaptação de Baixo Rank Decomposta por Peso (DoRA)

A DoRA, proposta por Liu et al. (2024), foi desenvolvida com foco em reduzir a lacuna de precisão entre as abordagens PEFT e o FFT. Estudos sobre a dinâmica de atualização de pesos revelaram que, na LoRA convencional, as variações de magnitude e de direção ocorrem de maneira acoplada e com correlação positiva, um comportamento que restringe a expressividade do aprendizado e difere substancialmente do FFT. No ajuste fino completo, o modelo é capaz de realizar ajustes sutis de direção simultaneamente a grandes alterações de magnitude, conferindo maior flexibilidade ao processo de aprendizagem (LIU et al., 2024).

Para contornar essa limitação, Liu et al. (2024) propuseram decompor explicitamente os pesos pré-treinados em dois componentes distintos: magnitude e direção. O vetor de magnitude é treinado diretamente, permitindo ajustes escalares independentes por camada. O componente direcional, de maior dimensionalidade, é otimizado de forma eficiente por matrizes de baixo rank típicas da LoRA. Essa separação permite que o modelo replique com maior fidelidade o padrão de atualização característico do FFT, sem incorrer nos custos computacionais associados ao ajuste fino completo.

Os resultados empíricos reportados por Liu et al. (2024) demonstraram que essa inovação melhora a estabilidade do treinamento e eleva significativamente a precisão em benchmarks de raciocínio lógico e instrução visual, aproximando o PEFT da qualidade do FFT sem introduzir sobrecarga adicional na fase de inferência. A DoRA superou consistentemente a LoRA padrão em todas as tarefas avaliadas pelos autores, consolidando-se como a abordagem mais indicada quando a prioridade é a maximização da capacidade expressiva do modelo adaptado.

2.3 Comparativo metodológico e diretrizes de aplicabilidade

A análise das três variantes revela perfis de desempenho distintos, que se adequam a necessidades operacionais igualmente distintas. Para sistematizar esse panorama, o Quadro 1 apresenta as principais características estruturais de cada técnica e as respectivas recomendações de aplicabilidade, com base nas análises de He et al. (2025) e nas métricas reportadas pelos autores originais.

Quadro 1: Comparação Metodológica das Variantes de Adaptação de Baixo Rank

Característica/ Método	LoRA (Hu et al., 2021)	QLoRA (Dettmers et al., 2023)	AdaLoRA (Toksoz; Isik, 2026)	DoRA (Liu et al., 2024)
Foco de Otimização	Redução básica de parâmetros	Extrema eficiência de memória	Eficiência alocativa de parâmetros	Maximização da capacidade expressiva
Estratégia de Rank	Fixo e uniforme	Fixo e uniforme	Dinâmico (SVD) adaptativo	Fixo (com decomposição estrutural)
Uso de VRAM	Moderado	Muito Baixo (Quantização 4-bit)	Moderado a Baixo	Moderado
Aproximação com o FFT	Baixa a Moderada	Moderada	Moderada a Alta	Muito Alta
Melhor Caso de Uso	Adaptações gerais sem restrições severas.	Uso em GPUs de consumo únicas (single-GPU).	Ajustes onde minimizar pesos ativos é estritamente essencial.	Tarefas complexas (ex: raciocínio matemático, instrução visual).

Fonte: Elaborado pelo autor (2026), adaptado de He et al. (2026)

O panorama delineado pelo Quadro 1 evidencia que a escolha do método constitui, essencialmente, uma função de compromisso (*trade-off*) entre as restrições de infraestrutura disponível e as exigências cognitivas da tarefa-alvo (HE et al., 2026). Em ambientes com *hardware* severamente restrito, a QLoRA permanece como a solução mais adequada, viabilizando operações com LLMs robustos em equipamentos de consumo. Quando o foco recai sobre a minimização do número de parâmetros ativos em arquiteturas específicas, a AdaLoRA fornece o mecanismo mais avançado de seleção de componentes críticos. Por sua vez, sempre que a prioridade for a extração máxima de capacidade expressiva e a minimização da lacuna em relação ao FFT, a DoRA firma-se como a abordagem mais promissora para a especialização e adaptação de domínio dos modelos.

Cabe ressaltar que as variantes não são mutuamente exclusivas em todos os cenários. He et al. (2026) apontam que a combinação de quantização com decomposição estrutural representa uma fronteira ativa de pesquisa, com potencial de reunir as vantagens da QLoRA e da DoRA em um único *framework*. Tais desenvolvimentos reforçam a relevância de compreender os fundamentos de cada variante antes de selecionar a estratégia mais adequada

ao contexto específico de aplicação. Apresentados os aspectos teóricos e analíticos, a seção seguinte descreve os procedimentos metodológicos que orientaram a condução desta revisão.

3. METODOLOGIA

Este estudo caracteriza-se como uma pesquisa exploratória fundamentada em revisão bibliográfica de caráter narrativo e análise comparativa da literatura científica recente. A opção pela revisão narrativa justifica-se pelo objetivo central do trabalho: em vez de mapear exaustivamente toda a produção sobre ajuste fino eficiente em parâmetros, buscou-se identificar, analisar criticamente e comparar as principais variantes da LoRA com base em suas inovações arquitetônicas e evidências empíricas de desempenho.

A coleta do material bibliográfico ocorreu por meio do levantamento de artigos seminais, anais de conferências de alto impacto no campo da inteligência artificial, como a International Conference on Learning Representations (ICLR) e a Neural Information Processing Systems (NeurIPS), além de publicações disponíveis no repositório arXiv, abrangendo o período de 2020 a 2026. A string de busca utilizada nas plataformas de busca acadêmica foi: ("LoRA" OR "Low-Rank Adaptation") AND ("LLM" OR "Large Language Model") AND ("fine-tuning" OR "PEFT").

Os critérios de inclusão priorizaram: (a) estudos empíricos que propuseram ou avaliaram as variantes LoRA, QLoRA, AdaLoRA e DoRA; (b) trabalhos que reportaram métricas objetivas de uso de VRAM, redução de parâmetros treináveis e desempenho em benchmarks de NLP; e (c) artigos publicados em periódicos com revisão por pares ou disponibilizados em repositórios de preprints de reconhecida relevância pela comunidade científica, como o arXiv.

Os critérios de exclusão abrangeram: (a) estudos sem resultados empíricos reportados; (b) publicações anteriores a 2020, período que precede a proposição formal da LoRA por Hu et al. (2021); (c) trabalhos que não avaliavam especificamente as variantes selecionadas; e (d) publicações em idiomas que não o inglês ou o português. O corpus final consistiu em oito obras selecionadas pela centralidade de sua contribuição ao tema. A análise dos dados fundamentou-se na extração de métricas de desempenho e eficiência para a construção do quadro comparativo qualitativo apresentado na seção anterior.

4. CONSIDERAÇÕES FINAIS

O presente estudo evidenciou que não existe uma técnica universalmente superior no ecossistema de otimização de LLMs, mas sim escolhas estratégicas determinadas pelas restrições de hardware e pelos requisitos de precisão de cada projeto. A análise comparativa demonstrou que, embora a LoRA original tenha estabelecido os fundamentos do ajuste fino eficiente em parâmetros, suas limitações expressivas, decorrentes da imposição de um rank fixo e uniforme, tornaram necessárias inovações estruturais que seguem trajetórias distintas e complementares.

A QLoRA mostrou-se indispensável para a democratização do acesso à inteligência artificial em contextos acadêmicos e de hardware de consumo, ao tornar viável o ajuste fino de modelos de grande escala em GPUs comerciais por meio da quantização em 4 bits. A DoRA representa o avanço mais expressivo em direção à qualidade do aprendizado profundo sem os custos do FFT, ao solucionar o problema do acoplamento entre magnitude e direção dos pesos por meio de uma decomposição estrutural inovadora (LIU et al., 2024). A AdaLoRA, embora teoricamente elegante em sua abordagem de alocação dinâmica de ranks via SVD, exige maior tempo de treinamento para consolidar suas vantagens práticas em relação às abordagens de rank fixo (ZHANG et al., 2023).

Do ponto de vista das contribuições teóricas, este artigo organiza e sistematiza as evidências dispersas na literatura sobre as três principais variantes da LoRA, propondo um quadro comparativo que pode servir de referência para pesquisadores e profissionais que buscam selecionar a abordagem mais adequada a seus contextos operacionais. Como limitação, destaca-se que a pesquisa fundamentou-se na compilação e análise crítica de literatura primária, sem a condução de experimentos empíricos próprios para validação cruzada dos benchmarks reportados pelos autores originais.

Para trabalhos futuros, recomenda-se a realização de testes laboratoriais unificados que submetam as três metodologias às mesmas tarefas e configurações de hardware, gerando dados comparativos primários que contribuam para o ecossistema nacional de pesquisa em inteligência artificial. Adicionalmente, a investigação de abordagens híbridas que combinem os mecanismos de quantização da QLoRA com a decomposição estrutural da DoRA representa uma direção promissora, alinhada às tendências mais recentes identificadas por He et al. (2026).

REFERÊNCIAS

AGHAJANYAN, Armen; GUPTA, Sonal; ZETTLEMOYER, Luke. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. arXiv, 2020. Disponível em: https://arxiv.org/abs/2012.13255. Acesso em: 13 mar. 2026.

DETTMERS, Tim; PAGNONI, Artidoro; HOLTZMAN, Ari; ZETTLEMOYER, Luke. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv, 2023. Disponível em: https://arxiv.org/abs/2305.14314. Acesso em: 22 mar. 2026.

HE, Haonan et al. A Unified Study of LoRA Variants: Taxonomy, Review, Codebase, and Empirical Evaluation. arXiv, 2026. Disponível em: https://arxiv.org/abs/2601.22708. Acesso em: 03 abr. 2026.

HU, Edward J. et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv, 2021. Disponível em: https://arxiv.org/abs/2106.09685v2. Acesso em: 05 abr. 2026.

LIU, Shih-Yang et al. DoRA: Weight-Decomposed Low-Rank Adaptation. arXiv, 2024. Disponível em: https://arxiv.org/abs/2402.09353. Acesso em: 17 abr. 2026.

ZHANG, Qingru et al. AdaLoRA: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. arXiv, 2023. Disponível em: https://arxiv.org/abs/2303.10512. Acesso em: 15 abr. 2026.

TOKSOZ, Seda Bayat; ISIK, Gultekin. A Comparative Evaluation of QLoRA and AdaLoRA for Parameter-Efficient Fine-Tuning of Large Language Models on Medical Textbook Question Answering. Artificial Intelligence in Applied Sciences, Igdir, v. 2, n. 1, p. 27-31, 2026. DOI: https://doi.org/10.69882/adba.ai.2026014.

ZENG, Yuchen; LEE, Kangwook. The Expressive Power of Low-Rank Adaptation. arXiv, 2023. Disponível em: https://arxiv.org/abs/2310.17513. Acesso em: 06 maio 2026.

Agradecimentos

Agradeço expressamente à minha professora e orientadora, Adriane Cavichiolli, por toda a dedicação, paciência e excelência no direcionamento desta pesquisa. Seus ensinamentos, suporte contínuo e orientação atenta foram pilares fundamentais para o desenvolvimento e a conclusão deste artigo.