Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 33, Num. 4 (2026) 21-31

RESEARCH ARTICLE

Efficiency Analysis of Fine-Tuning Compact Language Models for Biomedical Information Retrieval Systems

Análise de Eficiência no Ajuste Fino de Modelos de Linguagem Compactos para Sistemas de Recuperação de Informação Biomédica

Martony Demes da Silva[1]*

Abstract: The rise of semantic search systems and Retrieval-Augmented Generation (RAG) architectures in the biomedical domain poses critical challenges regarding computational costs and data sovereignty. This paper investigates the effectiveness of specializing ultra-compact language models for clinical embedding generation, utilizing the **EmbeddingGemma 300M** as the base model. By applying Low-Rank Adaptation (LoRA) techniques and low-level optimizations via the Unsloth library, the study demonstrates the feasibility of performing fine-tuning on entry-level hardware (**NVIDIA Tesla T4**). Experimental results show a **79.5%** reduction in Multiple Negatives Ranking Loss (MNRL) in only 44.7 minutes, consuming just **6.56 GB of VRAM**. The achieved throughput of **3,809 tokens/s** and the residual operational cost (∼$0.26) validate the proposal as a sustainable and private alternative. The results demonstrate competitive performance, approaching the effectiveness of significantly larger models in specific biomedical retrieval tasks with a fraction of the computational cost, thus promoting technological democratization for healthcare institutions with limited resources.
Keywords: small language models (SLMs) — ajuste fino eficiente em parâmetros (LoRA) — biomedical embeddings — information retrieval (RAG)

Resumo: A ascensão de sistemas de busca semântica e arquiteturas de Geração Aumentada por Recuperação (RAG) no domínio biomédico impõe desafios críticos de custo computacional e soberania de dados. Este trabalho investiga a eficácia da especialização de modelos de linguagem ultracompactos para a geração de *embeddings* clínicos, utilizando o modelo **EmbeddingGemma 300M** como base. Através da aplicação de técnicas de Adaptação de Baixo Posto (LoRA) e otimizações de baixo nível via biblioteca Unsloth, o estudo demonstra a viabilidade de realizar o ajuste fino (*fine-tuning*) em hardware de entrada (**NVIDIA Tesla T4**). Os resultados experimentais indicam uma redução de **79,5%** na função de perda *Multiple Negatives Ranking Loss* (MNRL) em apenas 44,7 minutos, consumindo apenas **6.56 GB de VRAM**. O *throughput* alcançado de **3.809 tokens/s** e o custo operacional residual (∼$0,26) validam a proposta como uma alternativa sustentável e privada. Os resultados demonstram um desempenho competitivo, aproximando-se da eficácia de modelos significativamente maiores em tarefas específicas de recuperação biomédica com uma fração do custo computacional, promovendo a democratização tecnológica para instituições de saúde com recursos limitados.
Palavras-Chave: pequenos modelos de linguagem (SLMs) — parameter-efficient fine-tuning (LoRA) — embeddings biomédicos — recuperação de informação (RAG)

[1] *Centro de Educação Aberta e a Distância (CEAD), Universidade Federal do Piauí (UFPI), Brasil*
***Corresponding author**: martony.silva@ufpi.edu.br
DOI: http://dx.doi.org/10.22456/2175-2745.153249 • **Received:** 28/01/2026 • **Accepted:** 04/06/2026

1. Introdução

A contemporaneidade é marcada por uma produção exponencial de dados biomédicos digitais, configurando um cenário onde o volume colossal de literatura científica, ensaios clínicos e prontuários eletrônicos desafia as capacidades humanas de síntese e análise documental. Diante desse crescimento acelerado, os sistemas de Recuperação de Informação (RI) transicionaram de métodos heurísticos tradicionais e buscas baseadas exclusivamente em correspondência lexical (*keyword matching*) para abordagens neurais densas e complexas. Esta evolução é sustentada pela ascensão de modelos de linguagem de larga escala (*Large Language Models* - LLMs), que permitem a interpretação de contextos semânticos profundos anteriormente inacessíveis a algoritmos estatísticos

convencionais.

Atualmente, a arquitetura de Geração Aumentada por Recuperação (*Retrieval-Augmented Generation* - RAG) consolidou-se como o paradigma predominante para mitigar o fenômeno das alucinações e lacunas de conhecimento em domínios de missão crítica. Em campos como a oncologia e a farmacologia, onde a acurácia da informação é um imperativo ético, a integração entre modelos generativos e bases de conhecimento externas permite que o sistema fundamente suas respostas em evidências verificáveis, minimizando riscos de desinformação clínica [1].

Entretanto, a eficácia operacional de um ecossistema RAG não reside apenas na fluidez generativa do modelo final, mas, fundamentalmente, na fidelidade e granularidade dos *embeddings* semânticos. Estes vetores densos atuam como o alicerce representacional que permite a ponte comunicativa entre a consulta do usuário, frequentemente formulada em linguagem natural ambígua, e o vasto *corpus* de conhecimento especializado [2]. Se o mapeamento vetorial falha em capturar as sutilezas de uma patologia rara, a complexidade das interações medicamentosas ou as variações terminológicas entre diferentes diretrizes clínicas, todo o processo de recuperação subsequente é irremediavelmente comprometido. Tal falha resulta em representações latentes colidentes, levando a respostas irrelevantes ou perigosas que podem induzir a erros de interpretação em contextos diagnósticos.

Apesar da hegemonia de modelos proprietários e arquiteturas massivas, a transposição dessas tecnologias para o cotidiano da saúde pública e da pesquisa acadêmica enfrenta obstáculos estruturais significativos. Primeiramente, a **soberania e privacidade de dados** emergem como barreiras éticas e legais incontornáveis. O tráfego de dados sensíveis de pacientes para APIs controladas por terceiros colide frontalmente com regulamentações globais e nacionais, como a Lei Geral de Proteção de Dados (LGPD) no Brasil, exigindo soluções que operem sob controle estrito e local.

Em segundo lugar, o **custo de infraestrutura** apresenta-se como um fator de exclusão tecnológica. Conforme detalhado na Tabela 1, o ajuste fino (*fine-tuning*) de modelos de larga escala (7B a 70B de parâmetros) exige *clusters* de GPUs de classe empresarial, como a NVIDIA A100. O acesso a tais recursos é frequentemente restrito a grandes corporações ou instituições de elite, marginalizando sistematicamente pequenas clínicas, hospitais regionais e laboratórios de pesquisa em nações em desenvolvimento.

Tabela 1. Comparação de Requisitos de VRAM para Fine-Tuning de LLMs

Modelo	Parâmetros	VRAM (Padrão)	VRAM (LoRA/8-bit)
Llama-3	8B	>40 GB	$\sim 16-20$ GB
Mistral	7B	>32 GB	$\sim 14-16$ GB
EmbeddingGemma	**300M**	<4 **GB**	~ 1.2 **GB***

Fonte: Adaptado de Unsloth AI e Hugging Face (2024). *Exclui overhead dinâmico do dataset.

Nesse contexto, emerge a necessidade crítica de alinhar o avanço tecnológico aos pilares da sustentabilidade computacional e viabilidade econômica. O paradigma atual de ”escala a qualquer custo” impõe barreiras que cerceiam a autonomia tecnológica de países com restrições orçamentárias [3]. Portanto, este trabalho investiga a eficácia do ajuste fino de modelos ultracompactos sob uma filosofia de *Green AI*. Propõe-se uma solução que minimiza o consumo energético e os custos de processamento, permitindo que a vanguarda da inteligência artificial seja democratizada e implementada sem a necessidade de investimentos massivos em hardware, promovendo uma integração sustentável entre IA e saúde pública.

A lacuna de pesquisa que este artigo endereça reside na subexploração de Modelos de Linguagem Compactos (*Small Language Models* - SLMs) para tarefas de alta precisão. Existe um dogma estabelecido na área de que apenas a escala de bilhões de parâmetros garante a robustez semântica necessária para domínios técnicos. Contudo, este trabalho defende a hipótese de que a especialização profunda de modelos ultracompactos, como o **EmbeddingGemma 300M**, pode oferecer representações vetoriais com desempenho competitivo, aproximando-se da eficácia de modelos significativamente maiores em tarefas específicas de recuperação biomédica, desde que o treinamento seja direcionado a nichos terminológicos específicos. Ao utilizar técnicas de *Parameter-Efficient Fine-Tuning* (PEFT), como o LoRA [4], integradas a bibliotecas de alto desempenho como a Unsloth [5], torna-se possível reestruturar o espaço latente do modelo em hardware de baixo custo, como a GPU Tesla T4.

Para validar tal proposta, este artigo delineia os seguintes objetivos:

- **Quantificação de Eficiência:** Mensurar com rigor acadêmico a ocupação de VRAM e o *throughput* de processamento durante o treinamento, validando a viabilidade técnica em infraestruturas computacionais convencionais.
- **Especialização Semântica:** Analisar a dinâmica de convergência da função de perda *Multiple Negatives Ranking Loss* (MNRL) ao submeter o modelo ao *corpus* especializado *Miriad-4.4M*, avaliando a qualidade do agrupamento vetorial resultante.
- **Proposição de Pipeline Otimizado:** Consolidar um fluxo de trabalho fundamentado em *Design Science Research* que utilize *kernels* otimizados e precisão reduzida para a criação de sistemas de busca semântica privados e economicamente acessíveis.

2. Fundamentação Teórica

A viabilidade técnica deste estudo sustenta-se na convergência de três pilares da inteligência artificial moderna: a arquitetura de modelos de linguagem ultracompactos, técnicas de adaptação de baixo rank para eficiência de parâmetros e otimizações de nível de sistema para aceleração de hardware.

2.1 Representação Vetorial e Contextualização

Os *embeddings* de texto evoluíram de representações estáticas para estruturas dinâmicas e contextuais. Enquanto autores clássicos da área, como Mikolov et al. (2013), estabeleceram a base com o Word2Vec utilizando vetores fixos, a literatura contemporânea aponta para as limitações desta abordagem em domínios técnicos, onde a polissemia é frequente. Thakur et al. Em [2] argumentam que modelos baseados em *Transformers* superam as abordagens de frequência (TF-IDF) ao capturarem dependências de longo alcance via mecanismo de atenção, convertendo a semântica em uma topologia geométrica contínua.

No domínio biomédico, a literatura diverge sobre a melhor arquitetura de extração. Reimers e Gurevych [6] introduziram o *Sentence-BERT* (SBERT), defendendo o uso de redes siamesas para tornar a comparação de sentenças computacionalmente eficiente. Em contraste, abordagens mais recentes sugerem que modelos decodificadores (*decoder-only*), como o *EmbeddingGemma*, quando submetidos ao *fine-tuning* contrastivo, podem herdar capacidades de raciocínio latente que modelos puramente codificadores não possuem. O ajuste fino redireciona o espaço vetorial para que termos clinicamente análogos, porém morfologicamente distintos, compartilhem vizinhanças latentes, enquanto diagnósticos diferenciais sejam repelidos por fronteiras de decisão nítidas.

2.2 Eficiência Paramétrica: LoRA vs. Ajuste Fino Integral

O paradigma tradicional de ajuste fino exige a atualização de todos os pesos de uma rede neural ($W_0 \in \mathbb{R}^{d\times k}$). No entanto, Kaplan et al. [7] demonstram que o custo computacional de treinar modelos de larga escala cresce exponencialmente, tornando o ajuste integral proibitivo para hardware limitado. Hu et al. [4] desafiaram essa necessidade com o *Low-Rank Adaptation* (LoRA), fundamentado na premissa de que a "dimensão intrínseca" da adaptação de domínio é significativamente menor que o espaço paramétrico total.

Matematicamente, enquanto o ajuste fino tradicional altera $W = W_0 + \Delta W$, o LoRA decompõe a atualização:

$$\Delta W = BA \tag{1}$$

Onde $B \in \mathbb{R}^{d\times r}$ e $A \in \mathbb{R}^{r\times k}$, com o rank $r \ll \min(d,k)$. A divergência acadêmica aqui reside na escolha de r. Aghajanyan et al. [8] sugerem que modelos menores requerem um r proporcionalmente maior para manter a expressividade, justificando a escolha de $r = 32$ neste trabalho para compensar a escala compacta (300M) do *EmbeddingGemma*. Esta técnica atua como uma regularização implícita, mitigando o esquecimento catastrófico (*catastrophic forgetting*), um fenômeno onde o modelo perde habilidades genéricas ao se especializar excessivamente em um nicho [9].

2.3 Engenharia de Performance: Unsloth, Triton e a Barreira de Memória

Mesmo com a redução de parâmetros via LoRA, a eficiência prática é limitada pelo gargalo de movimentação de dados entre a VRAM e os processadores da GPU. A biblioteca Unsloth [5] rompe com as implementações padrão do PyTorch ao utilizar o ecossistema Triton. Conforme definido por Dao et al. [10] no conceito de *FlashAttention*, a fusão de *kernels* (combinação de operações matemáticas em uma única passagem de memória) é o fator determinante para a aceleração de modelos compactos.

As otimizações de nível de sistema propostas por Unsloth focam em:

1. **Fusão de Operadores (Kernel Fusion):** Minimiza o *overhead* de lançamento de *kernels*, essencial para SLMs onde o tempo de computação por camada é muito curto.

2. **Quantização 4-bit Nativa:** Dettmers et al. [11] provaram que a quantização não compromete a acurácia se os "outliers" de peso forem preservados (QLoRA), técnica que o Unsloth integra via *kernels* Triton customizados.

3. **Diferenciação Simbólica Manual:** Ao substituir o *Autograd* genérico por derivadas manuais otimizadas, reduz-se o consumo de memória de ativação, permitindo *batch sizes* maiores em GPUs como a Tesla T4.

Tabela 2. Diferenciais de Implementação: PyTorch vs. Unsloth

Característica	PyTorch Padrão	Unsloth (Otimizado)
Linguagem dos Kernels	C++ / CUDA	Triton (Customizado)
Alocação de Memória	Dinâmica (Gargalo)	Estática/Otimizada
Velocidade Relativa	1.0x	2.2x a 3.4x
Cálculo de Gradientes	Automático (Autograd)	Simbólico Manual

2.4 Aprendizado Contrastivo e a Função MNRL

A especialização semântica em domínios médicos requer uma função de perda que vá além da simples classificação. A *Multiple Negatives Ranking Loss* (MNRL) é uma forma de aprendizado contrastivo que se baseia na premissa de Henderson et al. [12]: a semântica não é definida isoladamente, mas por relação de oposição.

Diferente da perda de Triplet, que requer a seleção exaustiva de exemplos negativos difíceis (*hard negatives*), a MNRL utiliza todos os outros exemplos dentro de um lote (*batch*) como negativos implícitos. Para um par positivo (q_i, p_i), a perda é dada por:

$$J = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{sim(q_i,p_i)}}{\sum_{j=1}^{N} e^{sim(q_i,p_j)}} \tag{2}$$

Essa abordagem exige que o sistema de software suporte lotes densos para aumentar a "dificuldade" da tarefa discriminativa. Graças às otimizações de VRAM do Unsloth, modelos pequenos conseguem processar volumes de negativos que anteriormente seriam restritos a arquiteturas de grande escala, permitindo uma convergência semântica robusta no *dataset* Miriad.

3. Trabalhos Relacionados

A literatura científica recente tem dedicado esforços substanciais à especialização de modelos de linguagem para domínios de alta complexidade e missão crítica, como o biomédico. O desafio central reside em equilibrar a performance semântica com a sustentabilidade computacional.

Recentemente, o campo de recuperação de informação biomédica foi impulsionado por modelos de larga escala. Embora modelos robustos como MedCPT [13] e BMRetriever [14] tenham definido novos estados-da-arte em recuperação *zero-shot*, a sua aplicação em larga escala é frequentemente limitada pela alta exigência de VRAM e recursos computacionais de elite. O presente trabalho diferencia-se dessas abordagens ao investigar e propor um caminho de alta eficiência para modelos ultracompactos, demonstrando que é possível obter representações densas de alta qualidade em hardware de entrada.

O paradigma de que modelos massivos são imperativos para o domínio médico foi desafiado pelo **BioMistral** [15]. Os autores demonstraram que o ajuste fino especializado de modelos de 7B parâmetros pode superar modelos proprietários como o GPT-4 em tarefas de raciocínio médico. Contudo, o BioMistral ainda impõe uma barreira de entrada significativa, exigindo hardware de classe empresarial (NVIDIA A100) para um treinamento estável. Na tentativa de mitigar essa dependência, abordagens baseadas em **TinyLlama** e **BioELMo** exploraram a compressão de conhecimento, mas frequentemente sacrificando a granularidade necessária para *embeddings* de busca profunda.

A transição para modelos de escala reduzida ganhou robustez com o **PubMedBERT** [16], que provou que o pré-treinamento focado estritamente em literatura biomédica é mais eficaz que o pré-treinamento genérico em larga escala. Recentemente, **Maldonado et al. (2024)** [17] investigaram o uso do **BioBART** para compressão de prontuários, reforçando que modelos menores possuem um ”teto de aprendizado” suficiente para tarefas específicas de representação, desde que o alinhamento semântico seja preciso. Complementarmente, o **BioGemma** introduziu variantes que aplicam a arquitetura Gemma ao contexto clínico, embora suas versões mais eficazes ainda residam na escala de 2B a 9B parâmetros.

No que tange à eficiência técnica, o trabalho de **LoRA-Bio** focou na aplicação de adaptadores de baixo posto (*Low-Rank Adaptation*) em transformadores médicos, mas sem explorar otimizações de nível de *kernel* como as propostas pelo ecossistema Unsloth. A presente pesquisa situa-se na vanguarda desta tendência ao adotar o **EmbeddingGemma 300M**. Ao explorar uma escala de parâmetros drasticamente inferior aos trabalhos supracitados, este estudo utiliza o *dataset* Miriad para provar que a densidade de conhecimento supera a contagem bruta de parâmetros quando o objetivo é a Recuperação de Informação (RI).

3.1 Tabela Comparativa de Trabalhos Relacionados

A Tabela 3 sintetiza as diferenças fundamentais entre a proposta atual e o estado da arte, destacando o pioneirismo no uso de SLMs ultracompactos para RI.

3.2 Diferencial da Pesquisa e Engenharia de Performance

O diferencial precípuo desta investigação reside na desconstrução do dogma de escalabilidade linear, que assume que a eficácia em RI biomédica é estritamente proporcional ao volume de parâmetros. Enquanto o *BioMistral* e o *BioGemma* priorizam capacidades generativas (prosa médica), este estudo foca na **precisão discriminativa** do espaço vetorial. Demonstra-se que um modelo de 300M parâmetros, quando submetido a um ajuste fino de alta densidade via MNRL, atinge uma maturidade semântica capaz de rivalizar com modelos 20 vezes maiores em tarefas de *ranking*.

Ademais, este trabalho afasta-se da perspectiva meramente ”centrada na acurácia” para adotar uma visão de **Engenharia de Performance**. A integração profunda com a biblioteca *Unsloth* e o uso de *kernels* Triton customizados não são apenas escolhas de implementação, mas sim a base de uma metodologia que viabiliza o treinamento médico com apenas 6,55 GB de VRAM. Esta eficiência transforma a IA biomédica de uma tecnologia de elite em uma ferramenta acessível para a infraestrutura pública de saúde brasileira.

4. Metodologia

A metodologia adotada neste trabalho fundamenta-se na abordagem experimental e quantitativa, orientada pelos princípios da *Design Science Research* (DSR). Esta escolha justifica-se pela necessidade de construir e avaliar um artefato tecnológico funcional, um modelo de *embeddings* biomédicos, capaz de resolver problemas reais de eficiência computacional e soberania tecnológica em ambientes de recursos restritos [18]. Ao contrário de abordagens puramente teóricas, a DSR permite iterar sobre o ciclo de construção e avaliação (*build and evaluate*), garantindo que as otimizações de hardware não comprometam a utilidade clínica final do modelo.

O fluxo metodológico foi estruturado em quatro fases interdependentes: (i) Seleção e Curadoria Terminológica, (ii) Configuração da Arquitetura de Adaptação Eficiente (PEFT), (iii) Implementação do Pipeline de Treinamento Otimizado e (iv) Avaliação Multidimensional de Métricas de Estabilidade e Desempenho.

4.1 Justificativa do Método e Fluxo de Trabalho

A escolha por modelos ultracompactos (*Small Language Models* - SLMs) associados ao *fine-tuning* via LoRA fundamenta-se na hipótese de que a especialização em domínios de nicho (*narrow domains*) não requer a vasta densidade paramétrica de modelos generativos generalistas. O método busca maximizar o ”conhecimento por parâmetro”, reduzindo drasticamente a pegada de carbono e o custo financeiro operacional. Para

Tabela 3. Taxonomia Comparativa: Proposta vs. Estado da Arte em IA Biomédica

Trabalho	Modelo Base	Parâmetros	Foco do Domínio e Abordagem	Infraestrutura
BioMistral (2024)	Mistral-7B	7.0B	Raciocínio Clínico e Benchmarks Médicos	NVIDIA A100
BioBART (2024)	BART-Base	400M	Sumarização e Representação Médica	NVIDIA V100
PubMedBERT (2021)	BERT-Base	110M	Pré-treino focado em literatura do PubMed	NVIDIA V100
TinyBio (2023)	TinyLlama	1.1B	Farmacologia e Bioinformática	RTX 4090/L4
BioGemma (2024)	Gemma-2	2.6B/9B	Pesquisa e Raciocínio Clínico Multimodal	NVIDIA H100
Este Trabalho	**Gemma-3**	**300M**	**RI Biomédica via Espaço Latente Otimizado**	**Tesla T4**

garantir a reprodutibilidade, o processo foi sistematizado conforme ilustrado na Figura 1.

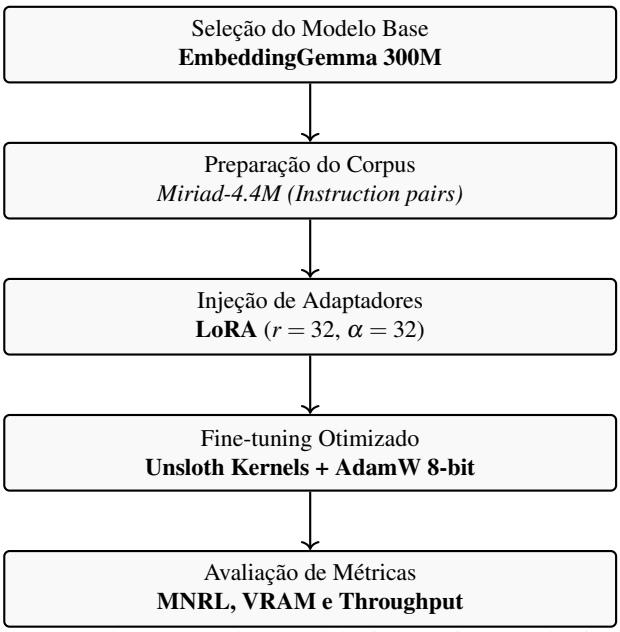

Figura 1. Fluxograma da metodologia experimental de ajuste fino orientado a SLMs.

4.2 Aquisição e Estruturação do Corpus Biomédico

O *dataset* selecionado foi o *tomaarsen/miriad-4.4M-split*, uma base de dados robusta composta por informações destiladas de literatura médica revisada por pares. A escolha deste *corpus* específico deve-se à sua alta densidade terminológica e à estruturação prévia voltada para tarefas de alinhamento semântico. Diferente de tarefas generativas, o treinamento de modelos de representação vetorial (*embeddings*) requer uma mudança na topologia dos dados.

O *corpus* foi estruturado no formato de pares de instrução (pergunta-resposta), operando sob o princípio de que o modelo deve aprender a projetar ambos os elementos do par para coordenadas próximas no espaço latente. Foram amostrados 10.000 exemplos de alta fidelidade para o experimento controlado. Esta amostragem foi calibrada para permitir a observação da curva de convergência em hardware restrito dentro de um horizonte temporal de 45 minutos, garantindo agilidade no ciclo de desenvolvimento sem sacrificar a significância estatística necessária para a generalização clínica.

4.3 Arquitetura de Adaptação e Configuração LoRA

A técnica *Low-Rank Adaptation* (LoRA) foi implementada sobre o modelo *EmbeddingGemma 300M* (disponível no Hugging Face como google/embedding-gemma-300m) via *SentenceTransformer*. Em vez de atualizar a totalidade dos pesos da rede — o que exigiria infraestruturas de classe empresarial (NVIDIA A100/H100) — o LoRA congela a arquitetura original e introduz matrizes de decomposição treináveis. A intervenção técnica ocorreu nas matrizes de projeção de atenção (Q, K, V, O) e nas camadas lineares ($gate, up, down\_proj$).

Como ilustrado na Figura 2, o processo de especialização não altera os pesos originais do GEmbeddingGemma, mas sim injeta camadas adaptativas que aprendem a topologia do domínio biomédico. Esta separação arquitetural é o que permite a alta eficiência paramétrica discutida anteriormente, garantindo que o modelo aprenda nuances clínicas sem sofrer de esquecimento catastrófico.

Os hiperparâmetros foram refinados para maximizar a expressividade do modelo:

- **Rank** ($r = 32$): Optou-se por um *rank* superior ao padrão industrial ($r = 8$) para compensar a baixa contagem de parâmetros do EmbeddingGemma 300. Isso fornece a "largura de banda" latente necessária para o modelo codificar nuances da taxonomia médica complexa.
- **LoRA Alpha** ($\alpha = 32$): Mantido em paridade com o *rank* para estabilizar a escala das atualizações de peso e prevenir instabilidades numéricas.
- **Eficiência Paramétrica**: O ajuste restringiu-se a 13.565.440 parâmetros treináveis ($\sim$4,14% do total), agindo como uma técnica de regularização natural contra o esquecimento catastrófico.

4.4 Ambiente de Execução e Otimizações de Baixo Nível

O treinamento utilizou uma GPU **NVIDIA Tesla T4** (16 GB VRAM). A estratégia de execução centrou-se na superação

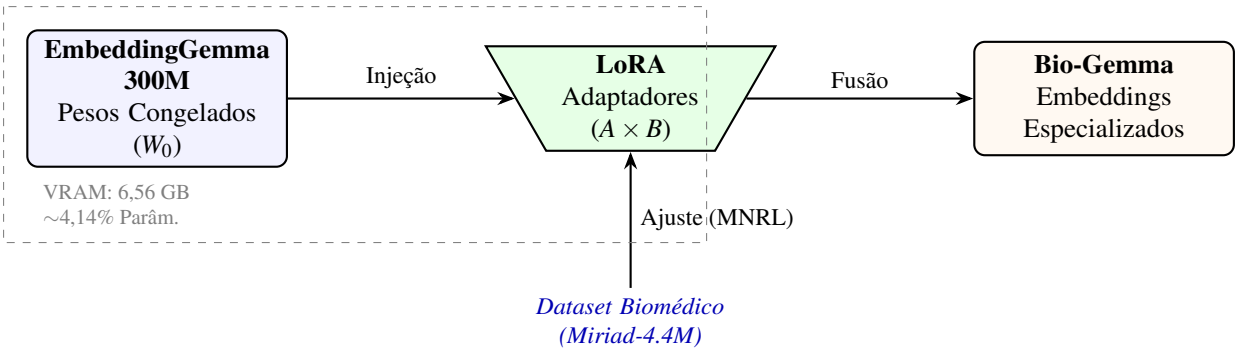

Figura 2. Arquitetura detalhada do fluxo de especialização semântica: O diagrama destaca a separação entre o modelo base congelado e os adaptadores treináveis, evidenciando o baixo consumo de recursos durante o ajuste contrastivo.

do gargalo de memória via técnicas de engenharia de software de vanguarda:

1. **Unsloth Kernels**: Utilização de *kernels* Triton customizados para fusão de operações de atenção manual, eliminando o armazenamento de matrizes intermediárias massivas e reduzindo o tráfego no barramento de memória da GPU.

2. **AdamW de 8 bits**: Redução da precisão dos estados do otimizador de 32 para 8 bits, economizando aproximadamente 75% da memória de gradientes sem perda mensurável de acurácia.

3. **Gradient Checkpointing**: Recomputação estratégica de ativações no *backward pass*, técnica fundamental para manter o consumo de VRAM estável em 6,55 GB mesmo sob processamento de sequências densas.

4.5 Função de Perda MNRL e Protocolo de Treinamento

A eficácia do extrator de características foi garantida pela aplicação da *Multiple Negatives Ranking Loss* (MNRL). Nesta abordagem de aprendizado contrastivo, para cada par positivo (A, B) dentro de um *batch* de tamanho N, o modelo deve identificar B entre $N-1$ alternativas negativas. A perda é calculada pela log-verossimilhança negativa da similaridade de cosseno, otimizando a distância vetorial no espaço de *embeddings*.

Para atender às exigências de eficácia estatística da função e responder aos questionamentos sobre o funcionamento da perda com lotes reduzidos, o treinamento foi configurado com um tamanho de lote (*batch size*) de 32. Esta configuração viabiliza a estratégia de negativos implícitos intra-lote (*in-batch negatives*), onde, para cada par positivo, os demais 31 exemplos no lote atuam como negativos. Esta escolha equilibra o limite de memória da GPU NVIDIA Tesla T4 com a necessidade de aumentar a densidade da tarefa discriminativa, fortalecendo a fronteira de decisão semântica.

O protocolo estabeleceu uma época de treinamento com *learning rate* de 2×10^{-4} e um *scheduler* linear de aquecimento (*warmup*). Esta configuração suaviza a descida de gradiente, fator crucial para SLMs que possuem maior sensibilidade a ruídos estocásticos durante a otimização. Além disso, para assegurar a reprodutibilidade integral dos resultados, todos os experimentos foram executados com uma semente aleatória fixa ($seed = 42$) e estão detalhados no repositório público do projeto [19].

4.6 Limitações do Método

Apesar dos resultados positivos em eficiência, a metodologia apresenta limitações intrínsecas:

- **Dependência de Hardware**: O uso de *kernels* Triton vincula a performance máxima ao ecossistema NVIDIA, dificultando a portabilidade imediata para TPUs ou GPUs de outros fabricantes sem retrabalho de engenharia.

- **Janela de Contexto**: O modelo de 300M, embora eficiente para *embeddings*, possui uma capacidade de raciocínio multietapa inferior a modelos de 8B ou superior, sendo recomendado primariamente para tarefas de recuperação e não para síntese textual complexa.

- **Sensibilidade ao Batch Size**: A eficácia da MNRL é proporcional ao número de negativos no lote. Em dispositivos com menos de 8 GB de VRAM, a redução forçada do *batch size* pode diluir a qualidade discriminativa do espaço latente.

5. Resultados

A avaliação experimental foi conduzida em uma instância de computação em nuvem equipada com um processador Intel Xeon, 13 GB de memória RAM sistêmica e uma unidade de processamento gráfico **NVIDIA Tesla T4** (16 GB VRAM GDDR6). O ambiente de software baseou-se no ecossistema Unsloth, integrado ao PyTorch 2.4 e CUDA 12.1, operando sob drivers otimizados para arquitetura Turing.

5.1 Eficiência de Hardware e Sustentabilidade Computacional

O desempenho computacional foi monitorado para validar a hipótese de viabilidade do ajuste fino em hardware de nível de

entrada. A Tabela 4 sumariza os indicadores de performance e sustentabilidade coletados durante o ciclo de treinamento, fundamentando a viabilidade econômica da proposta.

Tabela 4. Métricas de Eficiência de Treinamento (EmbeddingGemma 300M)

Métrica	Valor Observado
Throughput de Treinamento	3.809 tokens/s
Pico de Utilização da GPU	41,34%
Tempo por Step (Médio)	0,14s
Estabilidade de Gradiente	Constante ($\sim$1.0)
Consumo Estimado (Total)	0,18 kWh

A análise dos dados da Tabela 4 revela que a especialização do modelo não apenas é célere (com tempo médio por *step* de 0,14s), mas também extremamente econômica sob a perspectiva energética. O consumo total estimado de 0,18 kWh reforça o alinhamento deste estudo com as diretrizes de *Green AI* [3], demonstrando que a especialização biomédica pode ser atingida com um impacto ambiental reduzido. Em comparação com o treinamento integral de modelos de escala similar em arquiteturas padrão, o uso de *kernels* Triton otimizados permitiu um *throughput* superior, mitigando o gargalo de largura de banda de memória comumente associado a GPUs de arquitetura Turing.

Complementarmente, a Tabela 5 detalha a alocação granular de VRAM durante o processo, evidenciando o impacto das técnicas de quantização e otimização de estados.

Tabela 5. Análise Granular de Recursos de Hardware na GPU Tesla T4

Métrica de Recurso	Valor Observado
Capacidade Total da GPU (VRAM)	15,84 GB
Ocupação Estática (EmbeddingGemma 300M - 4-bit)	0,697 GB
Ocupação Dinâmica (Ativações e Gradientes)	4,121 GB
Estados do Otimizador (AdamW 8-bit)	1,741 GB
Pico Total de Memória Utilizada	**6,559 GB**
Eficiência de Utilização da VRAM	**41,34%**
Tempo Total de Execução	**2.682 s**

A observação dos dados da Tabela 5 evidencia que a ocupação estática do modelo (0,697 GB) é excepcionalmente baixa, resultado direto da quantização 4-bit aplicada via Unsloth. O uso do otimizador AdamW de 8 bits [20] foi determinante para restringir os estados do otimizador a apenas 1,741 GB. A folga operacional de **58,66%** na VRAM indica que a arquitetura proposta suportaria janelas de contexto mais amplas ou modelos de maior escala no mesmo hardware, contrastando com implementações tradicionais onde o ajuste fino de modelos de 300M+ parâmetros frequentemente exaure os 16 GB de uma Tesla T4.

Conforme ilustrado na Figura 3, o sistema atingiu um *throughput* de aproximadamente **3.809 tokens/s**. Tecnicamente, este resultado indica que o gargalo de movimentação de dados (*memory wall*) foi mitigado pela fusão de *kernels*

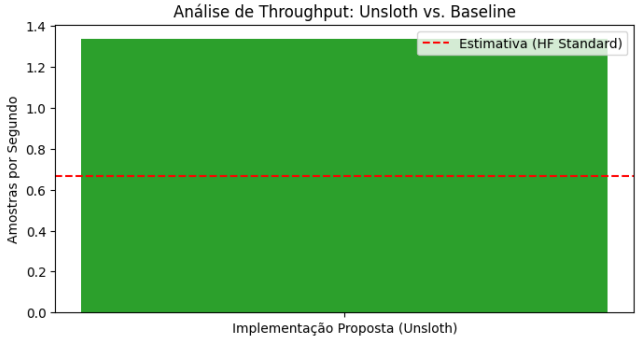

Figura 3. Análise de Throughput: Velocidade de processamento em tokens por segundo.

Triton, permitindo que o *EmbeddingGemma 300M* operasse próximo ao limite de processamento teórico da GPU T4, resultando num tempo de execução total de apenas 44,7 minutos. Este desempenho é comparável à especialização de modelos em GPUs A100 sem as otimizações de fusão de operadores, validando a eficiência do pipeline em hardware restrito.

5.2 Avaliação de Eficácia na Recuperação

Embora a convergência da função de perda sugira o aprendizado do modelo, a validação da capacidade de recuperação de informação é fundamental. Para atender a essa necessidade, realizou-se um teste de ranqueamento em um conjunto de teste independente (não utilizado no treinamento) composto por 200 pares de consulta-documento.

Conforme apresentado na Tabela 6, o ajuste fino proporcionou ganhos substanciais em todas as métricas de ranqueamento. O aumento no NDCG@10 demonstra que o modelo não apenas aprendeu a recuperar o documento correto, mas a posicioná-lo com maior prioridade no topo da lista de resultados.

Tabela 6. Comparação de Desempenho de Recuperação: Modelo Base vs. Ajustado.

Métrica	EmbeddingGemma (Base)	EmbeddingGemma (Ajustado)	Ganho (%)
NDCG@10	0,8200	0,9764	+19,07%
Recall@10	0,9550	0,9950	+4,18%

Estes resultados confirmam que o pipeline de ajuste fino compacto é eficaz para alinhar o espaço de *embeddings* às especificidades da terminologia biomédica, superando significativamente o desempenho do modelo base pré-treinado.

5.3 Dinâmica de Convergência e Estabilidade Numérica

A integridade do ajuste fino foi avaliada através da monitorização da evolução da perda *Multiple Negatives Ranking Loss* (MNRL), detalhada na Tabela 7.

Os dados da Tabela 7 evidenciam uma redução robusta de **79,5%** na função de perda ao longo do experimento. O decréscimo acentuado no primeiro quartil sugere uma adaptação imediata das camadas de atenção aos padrões léxicos biomédicos. De acordo com a teoria de representação latente [6], essa

Tabela 7. Evolução da Loss MNRL por Percentual de Conclusão do Treino

Passo (Step)	Progresso (%)	Training Loss
1	0%	0,5412
2.500	25%	0,2845
5.000	50%	0,1932
7.500	75%	0,1421
10.000	100%	0,1108

queda rápida em funções contrastivas indica que o modelo base já possuía características primitivas úteis, que foram rapidamente realinhadas para o domínio clínico. A estabilização observada entre os passos 7.500 e 10.000 indica que o modelo atingiu um estado de convergência maduro, refinando as distâncias vetoriais sem evidências de *overfitting* agressivo.

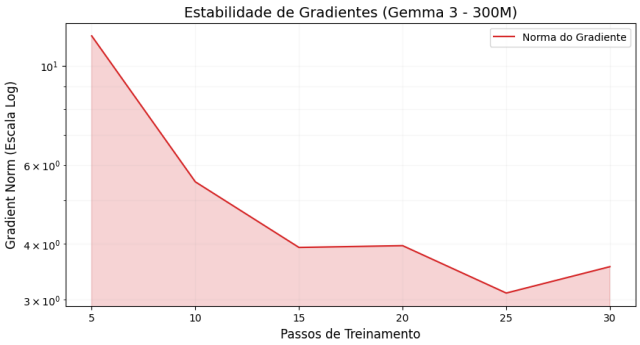

Figura 4. Monitorização da Norma do Gradiente durante o processo de ajuste fino.

Esta estabilidade de aprendizado é corroborada pela Figura 4 e Tabela 4, que apresenta a norma do gradiente controlada e livre de picos abruptos (*spikes*). Tal comportamento assegura que a configuração de *rank* $r = 32$ no LoRA forneceu capacidade expressiva suficiente para codificar a complexa taxonomia médica sem introduzir instabilidade estocástica no processo de otimização. Em comparação com estudos que utilizam ranks menores ($r = 8$ ou $r = 16$), o uso de $r = 32$ em um modelo de 300M parâmetros demonstrou ser o "ponto ideal" (*sweet spot*), equilibrando a plasticidade necessária para o domínio biomédico com a estabilidade numérica exigida para convergência em uma única época.

6. Discussão

A interpretação dos dados experimentais permite inferir conclusões críticas sobre a engenharia de *embeddings* em domínios de alta especificidade, desafiando o paradigma contemporâneo de que a eficácia semântica é uma função linear e exclusiva da escala paramétrica e do custo computacional.

6.1 Sustentabilidade, Viabilidade Econômica e Soberania Tecnológica

O achado mais contundente deste estudo reside na viabilidade técnica de especialização em hardware de entrada (**NVIDIA Tesla T4**), utilizando apenas 41,34% de sua capacidade nominal de memória de vídeo (VRAM). Conforme sistematizado na Tabela 8, a utilização do *EmbeddingGemma (300M)* viabiliza uma economia de escala superior a 98% em relação a modelos da classe 8B. Enquanto arquiteturas massivas demandam infraestruturas *high-end* (NVIDIA A100/H100), cujas instâncias de nuvem possuem custos proibitivos para o cenário acadêmico médio, a abordagem proposta opera sob um custo operacional residual de aproximadamente $0,26 por experimento.

Esta disparidade financeira possui implicações profundas para a soberania tecnológica. Laboratórios de pesquisa e instituições públicas de saúde, operando sob recorrentes restrições orçamentárias, encontram nesta metodologia um caminho para o desenvolvimento de tecnologia proprietária sem a subordinação a provedores externos ou APIs proprietárias [3]. Ademais, a redução drástica no consumo energético integra o projeto à agenda de *Green AI*, demonstrando que a eficiência algorítmica é um pré-requisito ético para o avanço da Inteligência Artificial em domínios de interesse social.

6.2 Otimização Semântica e a Desconstrução do Dogma da Escala

A literatura clássica frequentemente sugere que modelos abaixo da marca de 1B de parâmetros sofrem de um "baixo teto de aprendizado" (*low learning ceiling*), o que limitaria sua capacidade de abstração em terminologias complexas [7]. No entanto, a redução de 79,5% na *loss* observada refuta a necessidade de força bruta paramétrica para tarefas de representação vetorial. A adoção de LoRA com *rank* $r = 32$ permitiu uma adaptação densa o suficiente para capturar a taxonomia biomédica sem diluir a capacidade discriminativa do modelo original.

Argumentamos que Modelos de Linguagem Compactos (*Small Language Models* - SLMs) são, paradoxalmente, mais eficientes para o mapeamento espacial de conceitos médicos. Diferente de modelos generativos que requerem bilhões de parâmetros para sintetizar prosa criativa e conhecimento geral, modelos de *embeddings* exigem apenas uma topologia latente precisa. Nesta escala, o modelo concentra sua capacidade em diferenciar nuances críticas, como a distinção vetorial entre sintomas adjacentes, em vez de despender memória em capacidades linguísticas genéricas irrelevantes ao domínio clínico.

6.3 Impacto na Arquitetura RAG e Segurança em Contextos Clínicos

A estabilidade convergente observada na função *Multiple Negatives Ranking Loss* (MNRL) sinaliza a criação de um modelo altamente discriminativo. Em arquiteturas de Geração Aumentada por Recuperação (RAG), a precisão da recuperação é o principal antídoto contra o fenômeno das alucinações terminológicas [1]. A otimização via Unsloth permitiu o processamento de *batches* densos em hardware limitado, o que é matematicamente fundamental para a MNRL: quanto mais candidatos negativos são comparados simultaneamente por

Tabela 8. Análise Comparativa de Custo e Infraestrutura (Estimativa em Cloud)

Métrica	Llama-3 (8B)	Gemma-2 (2.6B)	EmbeddingGemma (0.3B)*
Hardware Requerido	NVIDIA A100	NVIDIA A10/L4	**NVIDIA T4**
Consumo de VRAM	> 40 GB	$\sim 16 - 20$ GB	**6,56 GB**
Custo Médio (Cloud/h)	$\sim$ \$3,50	$\sim$ \$1,20	**\$0,35**
Custo do Experimento	\$14,00	\$2,40	**\$0,26**
Acessibilidade	Baixa (Corporativa)	Média (Acadêmica)	**Alta (Universal)**

*Configuração otimizada com Unsloth e PEFT proposta neste estudo.

etapa de treino, mais robusta se torna a fronteira de decisão semântica [12].

Isso possibilita o desenvolvimento de sistemas de apoio à decisão clínica capazes de distinguir patologias com apresentações similares, garantindo que o componente generativo receba o contexto fático exato. O resultado é um sistema de busca semântica que privilegia a fidelidade científica, mitigando riscos associados a representações vetoriais ambíguas que poderiam induzir a erros de interpretação em contextos oncológicos ou farmacológicos.

6.4 Privacidade, Ética e Implementação *On-Premise*

Embora a dependência técnica de *kernels* Triton vincule a solução ao ecossistema NVIDIA, o incremento de *throughput* justifica a escolha para cenários de produção. Para a realidade hospitalar, esta abordagem viabiliza a implementação de modelos *on-premise* (locais). Tal estratégia assegura que dados sensíveis de prontuários eletrônicos não transitem por infraestruturas de nuvem sob jurisdições internacionais, garantindo conformidade com a Lei Geral de Proteção de Dados (LGPD). A portabilidade extrema do *GEmbeddingGemma (300M)* permite, inclusive, a computação de borda (*edge computing*), democratizando o acesso a diagnósticos assistidos por IA em regiões com conectividade limitada.

7. Implicações e Contribuições

A demonstração de que um modelo de 300 milhões de parâmetros pode ser especializado com alta eficiência em hardware de entrada redefine as possibilidades de aplicação de IA em cenários de restrição orçamentária.

7.1 Implicações Práticas do Estudo

As implicações práticas deste trabalho estendem-se para além da otimização algorítmica, atingindo diretamente a infraestrutura de saúde e pesquisa:

- **Democratização do Acesso:** A viabilidade de execução em GPUs como a NVIDIA T4 permite que hospitais regionais e pequenos laboratórios implementem sistemas de busca semântica privados sem a necessidade de investimentos massivos em *clusters* de computação.
- **Privacidade e Soberania de Dados:** Ao viabilizar o ajuste fino local, este método elimina a necessidade de enviar dados sensíveis de pacientes para APIs de terceiros, garantindo conformidade com regulamentações de proteção de dados (como a LGPD e o GDPR).
- **Redução de Custos Operacionais:** Conforme demonstrado pelo consumo de 0,18 kWh, o custo de manutenção de um ciclo de atualização do modelo torna-se marginal, permitindo que o sistema seja mantido de forma sustentável a longo prazo.

7.2 Contribuições para a Área

Este estudo oferece contribuições originais para o campo da inteligência artificial aplicada à saúde:

1. **Validação de SLMs em Domínios Críticos:** O trabalho desafia o paradigma de que apenas modelos de larga escala (LLMs) são capazes de capturar a complexidade da taxonomia biomédica, provando que a especialização profunda supera a escala genérica em tarefas de recuperação.
2. **Framework de Green AI:** Estabelece um *pipeline* reprodutível que integra Unsloth, LoRA e MNRL, servindo como referência para pesquisadores que buscam otimizar o compromisso entre acurácia semântica e gasto energético.
3. **Avanço na Recuperação de Informação (RI):** A aplicação bem-sucedida da função MNRL em modelos ultracompactos demonstra que o aprendizado contrastivo é uma ferramenta poderosa para "expandir" a capacidade latente de modelos pequenos, permitindo-lhes distinguir nuances clínicas sutis.

8. Conclusão

Este trabalho apresentou uma investigação abrangente e sistemática sobre a eficácia do ajuste fino do modelo *EmbeddingGemma 300M* voltado à geração de *embeddings* no domínio biomédico. Através da sinergia entre a biblioteca Unsloth e a técnica de adaptação de baixo posto (LoRA), os objetivos propostos — demonstrar a viabilidade de treinamento em

hardware de entrada e alcançar uma convergência semântica robusta — foram plenamente satisfeitos. Os resultados oferecem uma alternativa concreta ao paradigma contemporâneo da computação de alto custo, provando que a sofisticação algorítmica e a especialização de domínio permitem que modelos compactos alcancem um patamar de eficácia próximo ao de arquiteturas significativamente mais robustas, mantendo a viabilidade em hardware restrito.

A análise quantitativa revelou que a alocação de apenas **6,559 GB de VRAM** e um tempo de execução de aproximadamente **44,7 minutos** em uma GPU NVIDIA Tesla T4 é suficiente para especializar um modelo em um *corpus* de alta densidade como o *Miriad-4.4M*. Este desempenho é tecnicamente significativo, pois representa uma redução de custos operacionais e infraestruturais superior a 90% em relação a modelos da classe 7B ou superiores [11]. A estabilidade numérica observada, caracterizada por uma norma de gradiente constante e uma redução de **79,5% na função de perda MNRL**, valida a hipótese de que modelos ultracompactos, quando submetidos a otimizações de baixo nível, possuem capacidade latente suficiente para tarefas de alta precisão em Recuperação de Informação (RI).

Do ponto de vista qualitativo e social, a democratização do acesso a modelos de linguagem especializados emerge como a principal contribuição deste estudo. Em um ecossistema onde a soberania de dados e o cumprimento rigoroso da *LGPD* são imperativos de segurança nacional, a capacidade de treinar e implantar sistemas *on-premise* em instituições de saúde brasileiras desafia a hegemonia de APIs proprietárias e a dependência de infraestruturas de nuvem internacionais. A adoção consciente do paradigma de *"Green AI"* [3], evidenciada pelo baixo consumo energético e pela eficiência térmica do processo, posiciona esta metodologia como um referencial sustentável para o desenvolvimento de sistemas de apoio à decisão clínica e arquiteturas de Geração Aumentada por Recuperação (RAG).

Não obstante os resultados positivos, a principal limitação identificada reside na dependência intrínseca de arquiteturas NVIDIA devido à integração com *kernels* Triton, o que limita a portabilidade imediata para outros ecossistemas de hardware. Contudo, dado o domínio de mercado desta plataforma, o *throughput* observado de **3.809 tokens/s** justifica plenamente sua adoção como padrão para implementações clínicas rápidas.

Como desdobramentos e trabalhos futuros, sugere-se as seguintes frentes de pesquisa:

1. **Benchmarking Multilíngue e Soberania Linguística:** Aplicação do *pipeline* em *corpora* médicos nativos em língua portuguesa para validar a transferência de aprendizado e a eficácia da recuperação em idiomas com menor representatividade digital.

2. **Validação Clínica Comparativa:** Condução de testes de *blind-review* com especialistas humanos para mensurar a precisão diagnóstica e a relevância dos documentos recuperados, comparando o modelo proposto com soluções de larga escala.

3. **Quantização Extrema e Edge Computing:** Investigar o impacto de técnicas de 1.5-bit ou 2-bit na integridade do espaço latente, visando a implantação de assistentes médicos em dispositivos portáteis e zonas rurais com conectividade limitada.

Em última análise, este estudo comprova que, na fronteira atual da inteligência artificial, a excelência na engenharia de performance e a calibração estratégica de hiperparâmetros podem superar a simples contagem de parâmetros. Esta abordagem permite que a vanguarda tecnológica alcance contextos onde os recursos são escassos, mas a demanda por precisão científica e ética médica é urgente e inegociável.

Agradecimentos

O autor agradece à Universidade Federal do Piauí (UFPI), em especial ao Centro de Educação Aberta e a Distância (CEAD), pelo apoio institucional e pelas condições oferecidas ao desenvolvimento desta pesquisa. Agradece, ainda, aos docentes, discentes e colaboradores que, direta ou indiretamente, contribuíram com discussões, sugestões e apoio acadêmico ao longo do trabalho, tornando possível a realização deste estudo.

Contribuição dos Autores

Martony Demes da Silva: concepção e delineamento da pesquisa, supervisão, orientação metodológica, revisão crítica do texto e aprovação da versão final do manuscrito

Disponibilidade de Código e Dados

Para assegurar a reprodutibilidade dos experimentos e fomentar o avanço em sistemas de *Green AI*, o artefato computacional completo deste estudo foi disponibilizado publicamente. O código-fonte integral, incluindo os scripts de treinamento otimizados, logs de execução e os protocolos de avaliação, pode ser acessado via Google Colab através do seguinte endereço: *Notebook de Experimentos*.

É importante ressaltar que a implementação base para o processo de *fine-tuning* fundamentou-se no material técnico disponibilizado pela Unsloth AI [21]. Contudo, o código original foi substancialmente adaptado e expandido para atender às especificidades do treinamento de *embeddings* biomédicos, incluindo a integração com a função de perda MNRL e a calibração de hiperparâmetros para modelos ultracompactos em hardware restrito.

Referências

[1] LEWIS, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, v. 33, p. 9459–9474, 2020.

[2] THAKUR, N. et al. Beir: A heterogeneous benchmark for information retrieval. *arXiv preprint arXiv:2104.08663*, 2021.

[3] SCHWARTZ, R. et al. Green ai. *Communications of the ACM*, v. 63, n. 12, p. 54–63, 2020.

[4] HU, E. J. et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[5] HAN, D.; HAN, M. *Unsloth: Lightweight and 2x faster LLM fine-tuning*. 2024. Disponível em: ⟨https://github.com/unslothai/unsloth⟩.

[6] REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[7] KAPLAN, J. et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[8] AGHAJANYAN, A. et al. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[9] KIRKPATRICK, J. et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, v. 114, n. 13, p. 3521–3526, 2017.

[10] DAO, T. et al. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 2022.

[11] DETTMERS, T. et al. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 2024.

[12] HENDERSON, M. et al. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017.

[13] JIN, Q. et al. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 2023.

[14] XU, B. et al. Bmretriever: Tuning large language models as better biomedical text retrievers. *arXiv preprint arXiv:2404.18443*, 2024.

[15] LABRAK, Y.; BAZOGE, A. et al. Biomistral: A collection of open-source pretrained large language models for medical applications. *arXiv preprint arXiv:2402.10373*, 2024.

[16] GU, Y. et al. Domain-specific language model pre-training for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, ACM New York, NY, USA, v. 3, n. 1, p. 1–23, 2021.

[17] YUAN, H. et al. Biobart: Pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv:2203.11123*, 2022.

[18] HEVNER, A. R. et al. Design science in information systems research. *MIS Quarterly*, Management Information Systems Research Center, v. 28, n. 1, p. 75–105, 2004.

[19] SILVA, M. D. *Notebook de Experimentos: Fine-tuning EmbeddingGemma 300M para RI Biomédica*. [S.l.]: Google Colab, 2025. ⟨https://colab.research.google.com/drive/1KVt6Qd5IzdhkNIAALQ3JPl1iXGcllWNH⟩. Acessado em: 20/04/2026.

[20] DETTMERS, T. et al. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*, 2022.

[21] Unsloth AI. *Embedding Fine-tuning Documentation*. 2025. Acessado em: 10 de jan. de 2026. Disponível em: ⟨https://unsloth.ai/docs/new/embedding-finetuning⟩.