MIMIC-CXR-PT-BR e MedGemma ajustado para português brasileiro

Priscila Marques de Oliveira, Gustavo Freitas Alves

1 Dezembro 2025

Resumo

A tradução de relatórios radiológicos utilizando modelos de inteligência artificial para o português brasileiro visa ampliar o acesso ao conhecimento médico e auxiliar na saúde. Neste projeto buscamos efetuar a tradução de textos do dataset MIMIC-CXR do inglês para o português utilizando o modelo MedGemma 27B. Com estes dados, efetuamos o fine-tuning do modelo MedGemma 4B, utilizando o LoRA [1], uma técnica de adaptação de parâmetros eficiente, que nos permite diminuir custo computacional, sem perder desempenho. Os resultados mostram que apesar de não terem sido alcançados valores ótimos, o fine-tunning do MedGemma 4B para português é promissor e deve ser mais explorado.

1 Introdução

Traduzir textos médicos utilizando modelos de inteligência artificial é uma prática promissora e que tem ganhado muito espaço, especialmente nas áreas de exames de imagem. Contudo, existe uma baixa disponibilidade de dados médicos públicos em português e também de modelos de linguagem, se fazendo necessário a utilização da maioria em Inglês. Desta forma se torna importante que sejam desenvolvidas aplicações para o Português. Não se pode deixar de levar em conta que, relatórios de imagem, como os radiológicos, demandam uma precisão na avaliação e uso de terminologia médica, tornando a tradução automática um desafio maior. Já existem estudos na área biomédica [3] a respeito da utilização de LLMs com ajustes para domínios específicos. Estes estudos tem demonstrado um desempenho muito bom. Para treinar modelos grandes ou ajusta-los o custo é muito grande. Para tentar diminuir este problema técnicas como o LoRA [1] de adaptação de parâmetros permite ajustar o modelo, com redução do custo computacional e sem perda desempenho. Apesar disso, para a área médica ainda há escassez em estudos voltados para exames de imagens, principalmente em português, de tradução de relatórios. Por isso este trabalho propõe a tradução do dataset MIMIC-CXR para o português utilizando o MedGemma 27B, fine-tunning do MedGmma 4B utilizando LoRA com os dados obtidos através da tradução e a avaliação destes resultados.

2 Metodologia

Buscando aprimorar o desempenho do MedGemma4B em português brasileiro, foi realizado um ajuste fino do modelo utilizando laudos médicos do MIMIC-CXR traduzidos para o português. Esse conjunto será detalhado na Seção Conjunto de dados. Para a tradução dos laudos foi empregado o MedGemma27B, de modo a transferir características linguísticas do modelo maior para o modelo menor. A seguir são apresentados os modelos utilizados e seu papel no processo.

MedGemma4B

Para o ajuste fino e para as inferências do modelo final foi utilizada a versão multimodal google/medgemma-4b-it. Todas as execuções foram realizadas com quantização de 4 bits, reduzindo o consumo de memória sem comprometer a estabilidade do treinamento.
Nesta versão foi trabalhado algumas técnicas de prompt como Few Shot, onde era apresentado um exemplo de geração de laudo, e assim, requisitado a geração de texto a partir das imagens apresentadas. Além disso, verificamos, também, o desempenho para o modelo com Zero Shot.
Como prompt para as inferências foi utilizado a seguinte frase, "Dê o diagnóstico das imagens apresentadas".

MedGemma27B

O modelo google/medgemma-27b-text-it (texto-apenas) foi utilizado para a tradução das amostras de treino, validação e teste. Assim como no MedGemma4B, adotou-se quantização de 4 bits. Devido ao tamanho do modelo, foi empregada uma GPU NVIDIA A100 com 40 GB de VRAM. A tarefa consistiu em gerar traduções adequadas a partir de um prompt específico adaptado para laudos clínicos.
Prompt: "Você é um assistente médico que traduz relatórios de raio-x para o português. Traduza o relatório abaixo para o português."

MedGemma4B ajustado

Os hiperparâmetros utilizados no ajuste fino estão apresentados na Tabela 1. O LoRA foi aplicado em todas as camadas lineares do modelo. Devido às limitações de hardware, o treinamento foi realizado com batch size igual a 1 tanto no treino quanto na validação, limite de 2 imagens por amostra e acúmulo de gradiente de 16 passos, reduzindo significativamente o uso de VRAM.
As amostras traduzidas pelo MedGemma27B compuseram todo o conjunto de dados usado no ajuste fino. Como prompt para geração de texto foi utilizado do mesmo texto apresentao para as inferências, "Dê o diagnóstico das imagens apresentadas".

Parâmetro	Valor
α	16
Dropout	0.05
r	16

Tabela 1: Hiperparâmetros do LoRA.

Avaliação

A avaliação foi conduzida utilizando a métrica BERTScore em duas configurações:
(1) BERTimbau Base (neuralmind/bert-base-portuguese-cased), comparando o laudo de referência em português (traduzido pelo MedGemma27B) com o texto gerado pelo MedGemma4B;
(2) XLM-RoBERTa Large (xlm-roberta-large), comparando o laudo original em inglês (MIMIC-CXR) com o texto gerado em português, avaliando a preservação semântica entre idiomas.

3 Conjunto de dados

Neste projeto foi utilizado o MIMIC-Chest X-Ray (MIMIC-CXR), um conjunto composto por radiografias de tórax e respectivos laudos clínicos. O conjunto contém 377.110 imagens e 227.835 laudos, produzidos no Beth Israel Deaconess Medical Center (Boston, MA). As imagens possuem alta resolução, geralmente próximas de 2544×3056 pixels.
Os exames são organizados por paciente e estudo, e cada laudo possui anotações automáticas fornecidas pelo CheXpert e pelo NegBio. No presente estudo foram considerados apenas os rótulos do CheXpert, que abrangem 14 condições clínicas: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax e Support Devices.
A Tabela 2 apresenta a distribuição das classes entre casos confirmados e casos incertos.

Subconjunto de dados

Para viabilizar o ajuste fino do MedGemma4B sob limitações computacionais, foi elaborado um subconjunto reduzido do MIMIC-CXR. Esse subconjunto é composto por 960 estudos para treinamento, 473 para validação e 965 estudos para teste. Todas as imagens foram redimensionadas para 896×896 pixels, seguindo a resolução empregada pelo MedSigLip [2], o codificador de visão utilizado pelo MedGemma.

Rótulo	Positivo	Incerto
Atelectasis	45808	10327
Cardiomegaly	44845	6043
Consolidation	10778	4331
Edema	27018	13174
Enlarged Cardiomediastinum	7179	9375
Fracture	4390	555
Lung Lesion	6284	1141
Lung Opacity	51525	3831
No Finding	75455	0
Pleural Effusion	54300	5814
Pneumonia	16556	18291
Pneumothorax	10358	1134
Support Devices	66558	237

Tabela 2: Distribuição dos estudos conforme os rótulos do CheXpert. A coluna "Rótulo" representa a patologia; "Positivo" indica a quantidade de estudos onde a condição foi identificada; e "Incerto" representa os casos nos quais o modelo automático não pôde determinar com segurança a presença da condição.

4 Experimentos

Nossos experimentos consistem na aplicação da técnica LoRA [1] e nas inferências do modelo MedGemma4B utilizando das técnicas de prompt Zero Shot e Few Shot.

Treinamento do ajuste fino MedGemma4B

O ajuste fino do MedGemma4B foi realizado ao longo de 4 épocas, resultando em um tempo total de execução de 13 horas e 37 minutos. Os resultados de desempenho são apresentados nas Figuras 1, 2 e 3.
A partir da curva de loss de validação (Figura 1), observa-se um decaimento consistente ao longo do treinamento, com mínimo localizado em torno do passo 120, onde o valor atinge 0.9206. A métrica de Entropia (Figura 2) apresenta comportamento semelhante, com redução contínua durante o processo e valor mínimo de 0.4062. Já a acurácia média de tokens (Figura 3) demonstra crescimento até estabilizar próximo ao passo 180, atingindo o valor máximo de 0.7911.
Esses resultados indicam que a execução de 4 épocas foi além do necessário para esse conjunto de dados. As métricas convergem majoritariamente até o final da segunda época, sugerindo que treinos mais longos não trariam ganhos adicionais significativos.

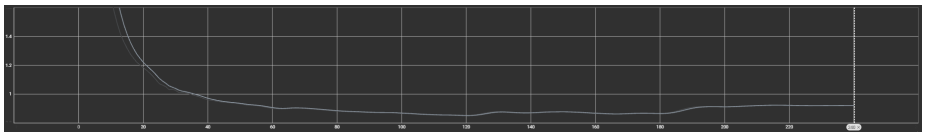

Figura 1: Evolução da loss de validação durante o ajuste fino do MedGemma4B. O valor mínimo é atingido próximo ao passo 120.

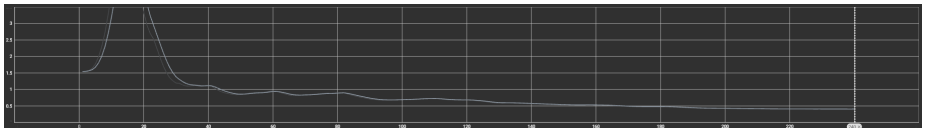

Figura 2: Evolução da entropia de validação durante o ajuste fino do MedGemma4B. O valor mínimo ocorre ao final do processo, indicando tendência de queda contínua caso mais épocas fossem treinadas.

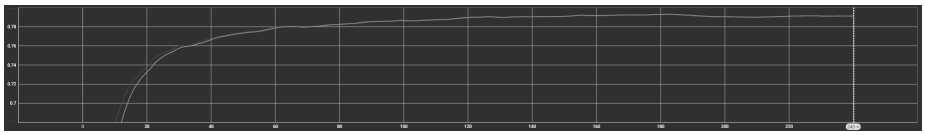

Figura 3: Evolução da acurácia média de tokens na validação durante o ajuste fino do MedGemma4B. O valor máximo é observado próximo ao passo 180.

ZeroShot x FewShot x LoRA

Considerando as três abordagens, temos os resultados apresentados nas Figuras 4, 5, 6, 7, 8 e 9.

Ao analisarmos as métricas obtidas com o modelo XLM-RoBERTa, observa-se que este apresenta melhores valores de precisão (4), recall (5) e f1-score (6) nos cenários de Few-Shot e LoRA, quando comparados ao caso Zero-Shot. Isso pode ser explicado pelo fato de que, no Few-Shot, fornecemos exemplos do formato textual esperado, o que tende a guiar a geração para tokens linguisticamente semelhantes. Para o LoRA, uma explicação similar se aplica, já que o modelo foi ajustado com relatórios do conjunto de dados MIMIC-CXR em português brasileiro.

Um ponto interessante é que as métricas do LoRA apresentam valores máximos e mínimos mais extremos, indicando maior variação ao longo do conjunto de teste. Isso sugere uma instabilidade maior na geração dos textos quando comparada aos demais cenários..

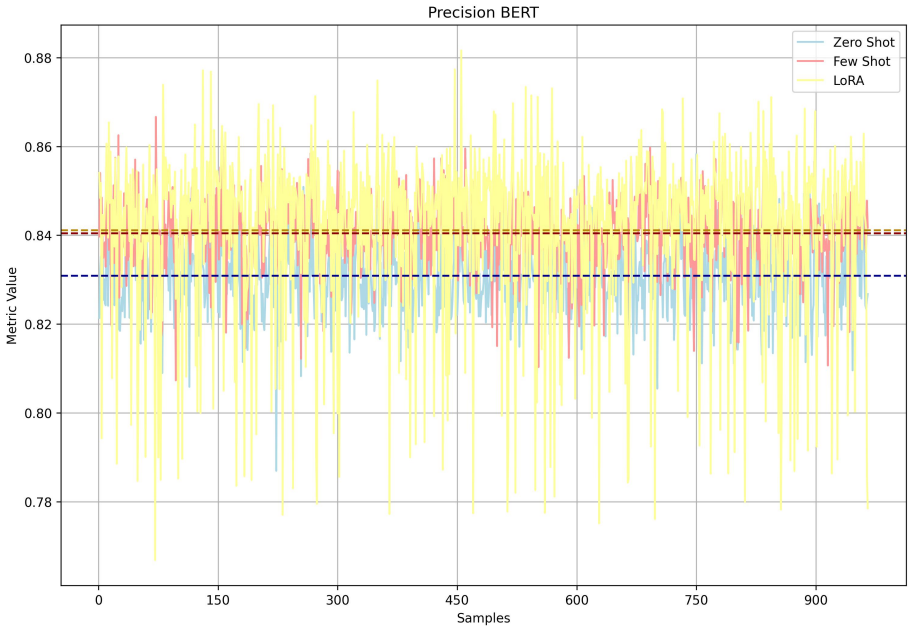

Figura 4: Métrica de precisão durante inferência com subconjunto de teste e utilizando modelo XLM-RoBERTa no BERTScore.

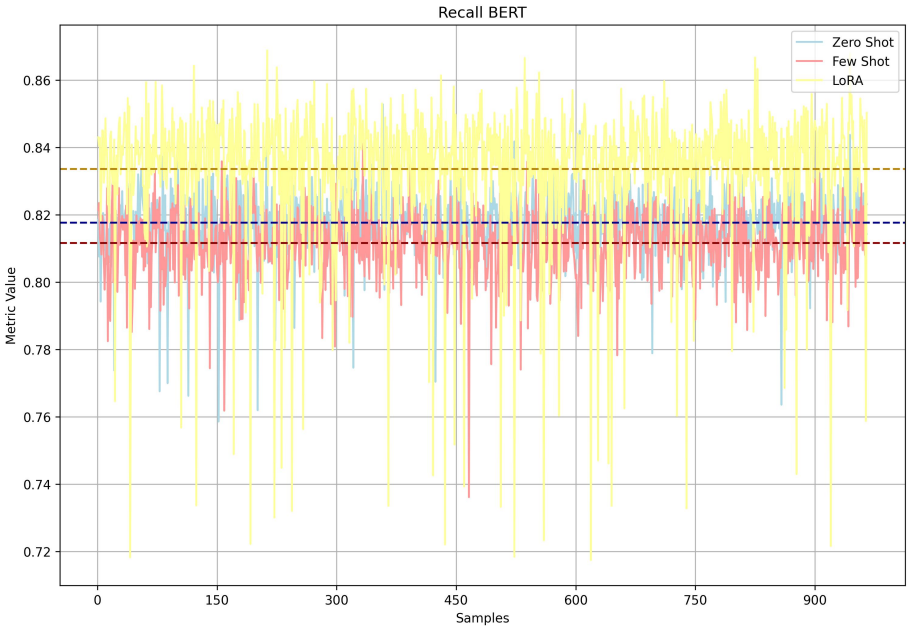

Figura 5: Métrica de recall durante inferência com subconjunto de teste e utilizando modelo XLM-RoBERTa no BERTScore.

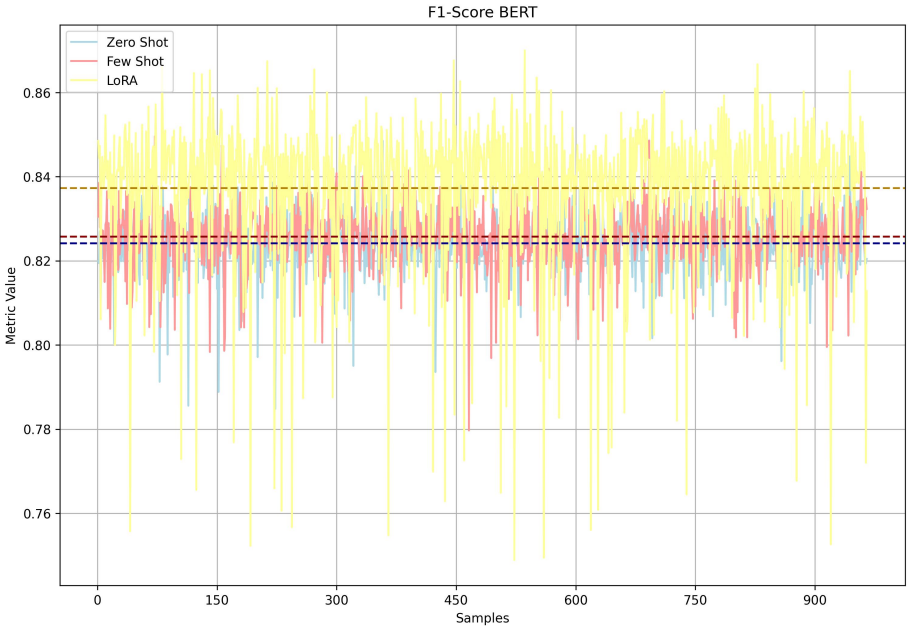

Figura 6: Métrica de f1-score durante inferência com subconjunto de teste e utilizando modelo XLM-RoBERTa no BERTScore.

Ao observar os resultados com o modelo BERTimbau, novamente verificamos uma melhora nos cenários Few-Shot e LoRA em relação ao Zero-Shot, tanto para precisão (7), quanto para recall (8) e f1-score (9). No entanto, nota-se que os valores mínimos obtidos pelo LoRA são menores que os demais, assim como seus valores máximos. Esse comportamento reforça a evidência de uma maior instabilidade na geração dos textos ajustados via LoRA.

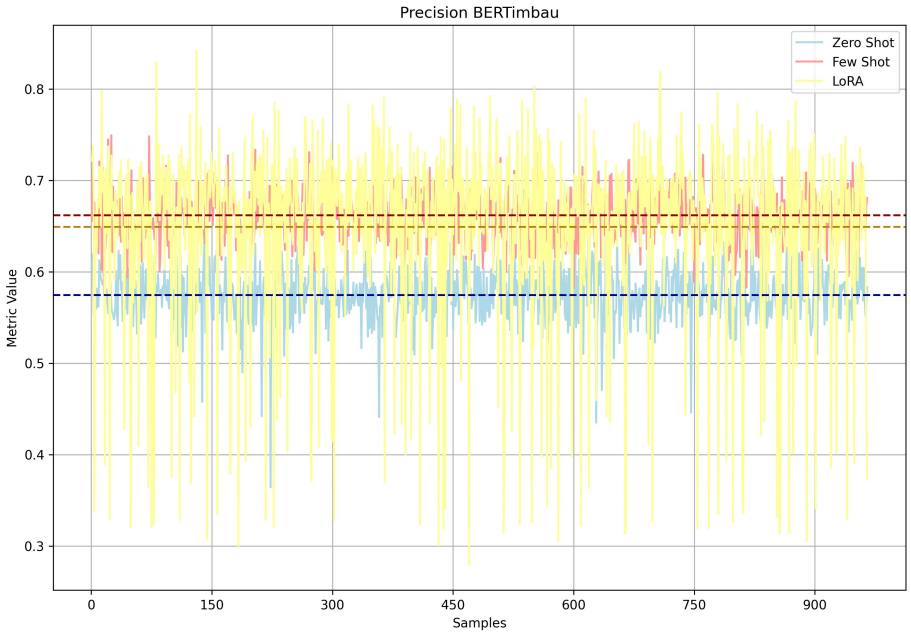

Figura 7: Métrica de precisão durante inferência com subconjunto de teste e utilizando modelo BERTimbau no BERTScore.

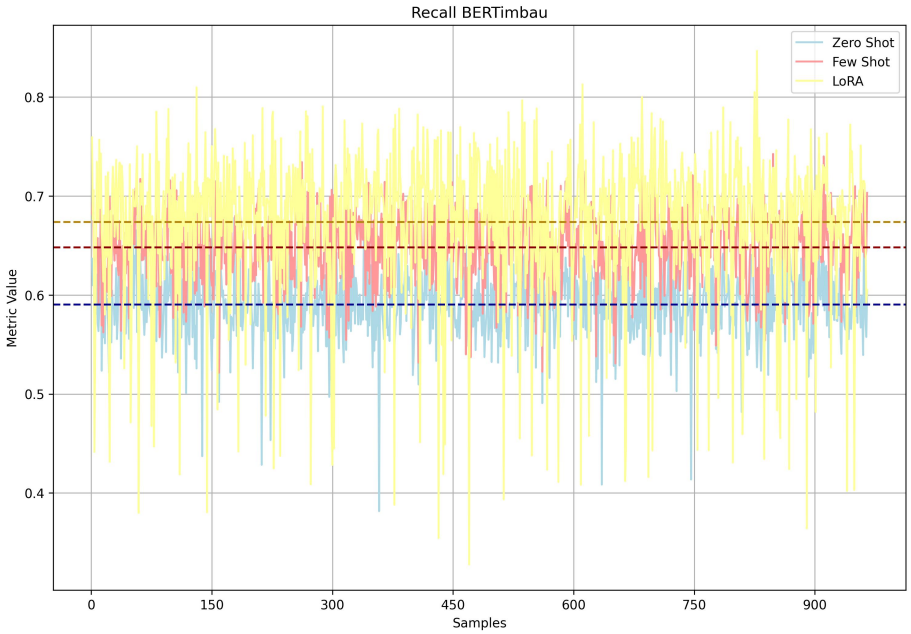

Figura 8: Métrica de recall durante inferência com subconjunto de teste e utilizando modelo BERTimbau no BERTScore.

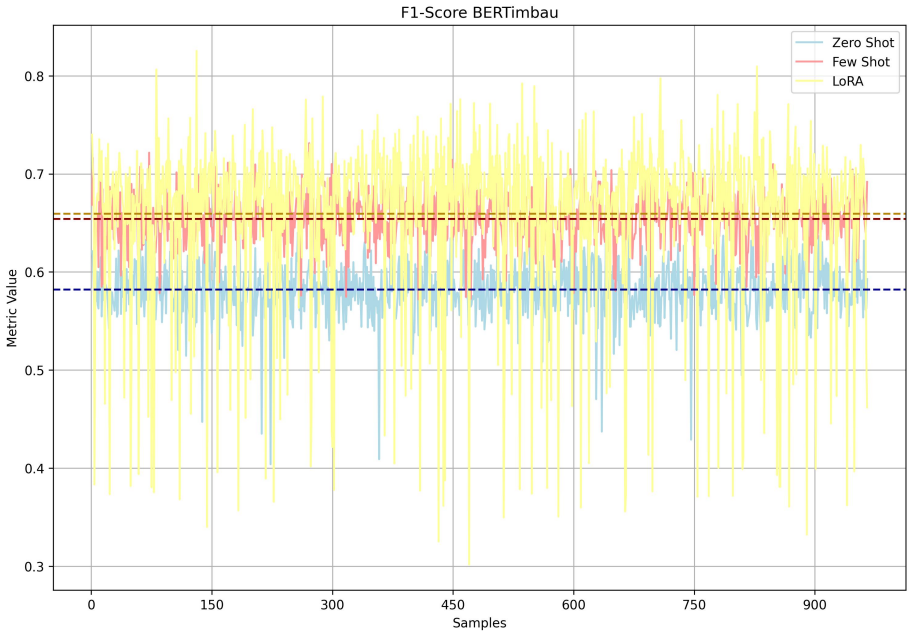

Figura 9: Métrica de f1-score durante inferência com subconjunto de teste e utilizando modelo BERTimbau no BERTScore.

5 Conclusão

Os experimentos demonstraram que o ajuste fino do MedGemma4B converge rapidamente, com métricas de validação estabilizando-se ainda na segunda época. Do ponto de vista da qualidade textual, observou-se que as abordagens Few-Shot e LoRA superam o Zero-Shot tanto nas avaliações com XLM-RoBERTa quanto com BERTimbau, indicando que fornecer exemplos ou realizar especialização do modelo favorece a geração de relatórios clínicos mais próximos da referência.

Entretanto, os resultados também evidenciam uma maior variabilidade no desempenho do modelo ajustado com LoRA. As amplitudes maiores entre mínimos e máximos nas métricas de precisão, recall e f1-score sugerem que o ajuste fino com conjunto reduzido, ainda que eficaz, introduz instabilidade na geração textual. Isso se deve ao conjunto de dados utilizados para o ajuste.

De modo geral, este estudo confirma o potencial do MedGemma4B, mesmo em sua versão compacta, para geração de laudos radiológicos em português, especialmente quando apoiado por técnicas de prompting ou ajuste leve de parâmetros. Os resultados evidenciam que modelos multimodais de médio porte podem ser adaptados com sucesso para domínios clínicos específicos, mesmo sob restrições computacionais significativas.

6 Trabalho futuro

Como continuidade deste estudo, alguns pontos podem ser melhor exploradas. O primeiro é a ampliação do conjunto de dados utilizado no ajuste fino do MedGemma4B, permitindo que o modelo se especialize melhor na geração de textos médicos em português brasileiro. Também é recomendável incorporar métricas adicionais, de modo a possibilitar comparações mais abrangentes entre diferentes estratégias de aprimoramento do modelo.
Outro ponto promissor envolve a investigação mais profunda de técnicas de prompting, já que ajustes na formulação das instruções podem impactar significativamente a qualidade das respostas. Da mesma forma, vale analisar com maior detalhe o comportamento do LoRA, explorando diferentes configurações de hiperparâmetros para otimizar seu desempenho.
Por fim, um aspecto essencial para a consolidação do método é a realização de uma avaliação humana especializada. A participação de profissionais de saúde permitirá validar a qualidade e a utilidade clínica dos laudos gerados, fornecendo uma análise que métricas automáticas não conseguem capturar.

Referências

[1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[2] Google Research. Med-gemma: A family of medical vision-language models. *arXiv preprint arXiv:2507.05201*, 2025.

[3] Hieu Tran, Zhen Yang, Ziyu Yao, and Hong Yu. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *arXiv preprint arXiv:2310.19975*, 2023.