# Synthetic Data Generation for Depression Detection via Parameter-Efficient Fine-Tuning of Large Language Model

**Carlos Henrique B. M. Leão**
UFMG - DCC
Belo Horizonte
chbmleao@ufmg.br

**Gabriel Bifano Freddi**
UFMG - DCC
Belo Horizonte
gabrielfreddi@info.grad.ufmg.br

**Tarcízio Augusto Santos Lafaiete**
UFMG - DCC
Belo Horizonte
tarcizio-augusto@ufmg.br

## Abstract

Este trabalho investiga a capacidade de um modelo de linguagem de porte moderado (1.1B parâmetros) em gerar dados sintéticos úteis para a detecção de discurso depressivo, enfrentando a limitação de bases clínicas pequenas e desbalanceadas, como o DAIC-WOZ. Após ajuste fino supervisionado do TinyLLaMA sobre 5.953 enunciados rotulados, utilizamos três checkpoints de treinamento para produzir amostras sintéticas condicionadas ao rótulo depressivo, visando balancear a base original. A utilidade dos dados gerados foi avaliada por meio do desempenho de um classificador treinado com dados reais, com métricas como acurácia e F1-score, e uma avaliação mais limitada feita utilizando uma LLM mais robusta como árbitro. Os resultados visam esclarecer se modelos compactos podem contribuir de forma eficaz para a expansão controlada de bases sensíveis em saúde mental, oferecendo uma alternativa viável para ambientes clínicos com restrições computacionais e limitações de coleta.

## 1 Introdução

A depressão é uma condição clínica grave e apresenta elevada prevalência mundial. Estimativas recentes da Organização Mundial da Saúde indicam que mais de 280 milhões de pessoas vivem com depressão, representando aproximadamente 3,8% da população global, incluindo 5,0% dos adultos e 5,7% das pessoas com 60 anos ou mais [11]. Dados do National Institute of Mental Health mostram que, apenas nos Estados Unidos, cerca de 21 milhões de adultos (equivalente a 8,3% da população) apresentaram ao menos um episódio depressivo maior em 2021 [10]. Informações epidemiológicas do CDC reforçam a expansão contínua da prevalência, apontando tendência de aumento particularmente entre jovens adultos [6]. Esses números evidenciam a necessidade de métodos complementares que ofereçam suporte ao diagnóstico clínico.

Apesar da relevância do problema, bases de dados rotuladas para diagnóstico assistido de depressão são escassas. O conjunto DAIC-WOZ, um dos mais utilizados, possui apenas algumas centenas de entrevistas, com limitações relacionadas à predominância de falantes de inglês, ao viés demográfico das coletas e, principalmente, ao tamanho reduzido e ao desbalanceamento entre classes (depressivo e não depressivo) [5]. Estudos de revisão mostram que bases para saúde mental apresentam fragmentação significativa, barreiras éticas de acesso, alto custo de anotação clínica e reduzida diversidade

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

populacional [2]. Mesmo trabalhos recentes em detecção de depressão por fala, como os apresentados no Interspeech 2025, destacam explicitamente que a pequena escala dos dados limita a robustez dos modelos e restringe sua capacidade de generalização [7]. Assim, a construção de bases maiores é necessária para apoiar psiquiatras e profissionais de saúde, possibilitando modelos mais sensíveis e com maior cobertura de padrões linguísticos.

Como a obtenção de novos dados clínicos depende de longos processos de coleta e aprovação ética, estratégias de aumento de dados tornam-se particularmente relevantes. Técnicas tradicionais utilizam métodos como back-translation, transformação semântica controlada e manipulação acústica, mas sua eficácia é limitada quando a base original é pequena ou heterogênea [1]. A introdução de modelos de linguagem de grande porte ampliou consideravelmente o repertório de técnicas disponíveis. Pesquisas recentes mostram que LLMs podem gerar exemplos sintéticos de alta variabilidade linguística, mantendo coerência com rótulos clínicos específicos [4, 12]. Esses modelos podem produzir paráfrases, novos enunciados condicionados a sintomas e até diálogos completos, sendo promissores para enfrentar a escassez de dados.

Ainda assim, o uso de dados sintéticos para tarefas clínicas permanece uma questão em aberto. Estudos na área de saúde mental indicam que dados gerados artificialmente podem melhorar o desempenho de classificadores quando utilizados de maneira controlada, mas também podem introduzir vieses ou inflar artificialmente métricas se não forem avaliados de forma rigorosa [8]. Trabalhos recentes também ressaltam que o comportamento de LLMs varia substancialmente conforme o tamanho do modelo e a quantidade de exemplos usados na fase de condicionamento [9]. Dessa forma, investigar modelos menores, da ordem de 1 bilhão de parâmetros, é essencial para aplicações práticas em ambientes clínicos, especialmente aqueles que exigem execução local, maior privacidade e restrição de recursos computacionais.

O objetivo deste trabalho é gerar dados sintéticos de alta qualidade que possam aumentar a robustez e a precisão de modelos baseados em dados. Para isso, utilizamos uma LLM pré-treinada com aproximadamente 1 bilhão de parâmetros e realizamos seu finetuning utilizando o conjunto de treinamento do DAIC-WOZ, composto por 5.953 pares de perguntas e respostas. Em seguida, avaliamos os dados sintéticos gerados — e, por consequência, o desempenho do modelo ajustado — por meio de métodos qualitativos e quantitativos.

## 2 Metodologia

A metodologia adotada neste estudo foi estruturada para avaliar a capacidade de um modelo de linguagem de porte moderado em gerar dados sintéticos úteis para a detecção de discurso depressivo. Esta seção descreve a arquitetura utilizada, o processo de ajuste fino, a estratégia de aumento de dados e o protocolo de avaliação aplicado.

Adotou-se a arquitetura *TinyLLaMA*, um modelo da família LLaMA contendo aproximadamente 1 bilhão de parâmetros. Essa escolha se justifica por suas características de leveza computacional, permitindo experimentação em ambientes com recursos restritos, além de possibilitar investigar a capacidade de modelos de menor escala na geração de dados sensíveis em contexto clínico.

O modelo foi submetido a *Supervised Fine-Tuning* utilizando um conjunto de 5.953 enunciados textuais rotulados do conjunto de treinamento. A distribuição das classes nesse conjunto consiste em 28,7% de enunciados depressivos e 71,3% de enunciados não depressivos, refletindo um cenário realista e significativamente desbalanceado. Adicionalmente, foram utilizados 1.785 exemplos para validação e 2.588 para teste, totalizando 10.326 pares pergunta-resposta processados.

Cada exemplo de treinamento foi formatado incluindo um contexto de até 3 pares pergunta-resposta anteriores do mesmo participante, permitindo que o modelo capture a continuidade conversacional. Além disso, cada prompt incluiu o escore PHQ-8 do participante (variando de 0 a 24), fornecendo informação contextual sobre a severidade dos sintomas depressivos.

Para viabilizar o ajuste fino com recursos computacionais limitados, foi empregada a técnica *Low-Rank Adaptation* (LoRA), que permite treinar apenas uma fração reduzida dos parâmetros do modelo. Especificamente, foram adaptados os módulos de projeção de consulta (*query*) e valor (*value*) das camadas de atenção. Esta configuração resultou em aproximadamente 1,1 milhão de parâmetros treináveis (0,1% do total), mantendo a eficiência computacional enquanto permite adaptação específica ao domínio.

Durante o processo de ajuste fino foram obtidos três *checkpoints* distintos. Cada um deles representa um estágio diferente de convergência do treinamento, o que permite avaliar como a progressão do ajuste influencia a qualidade da geração sintética. Esses três modelos resultantes são avaliados separadamente ao longo dos experimentos. A estratégia de avaliação durante o treinamento foi configurada para executar a cada 200 passos no conjunto de validação, permitindo monitoramento contínuo da métrica *loss* para a seleção do melhor modelo baseado neste critério.

Após o ajuste fino, cada modelo derivado dos três *checkpoints* foi utilizado para gerar novos enunciados sintéticos condicionados ao rótulo depressivo. Para cada versão do modelo, o objetivo foi duplicar o número de amostras depressivas originais, produzindo uma base de dados balanceada entre as classes.

Assim, para cada *checkpoint*, formou-se uma nova base expandida composta por:

- enunciados não depressivos reais (mantidos inalterados),
- enunciados depressivos reais (mantidos inalterados),
- enunciados depressivos sintéticos gerados pela LLM correspondente.

A utilidade dos dados sintéticos foi avaliada por meio de um classificador independente de discurso depressivo, treinado separadamente em cada uma das bases balanceadas e também com o conjunto de dados original. A arquitetura do classificador permaneceu fixa durante todo o experimento, garantindo que qualquer variação de desempenho decorresse exclusivamente da qualidade dos dados fornecidos. Para cada base gerada, o desempenho do classificador foi mensurado pela métrica *F1-score*, apropriada para cenários originalmente desbalanceados e para tarefas binárias envolvendo dados sensíveis.

Também foi feito um *score* de 0 a 10, com um número limitado de amostras sintéticas, utilizando o modelo do *Chat-GPT 4*, que é uma LLM robusta que se encontra no estado da arte. Foram selecionadas 20 amostras por modelo de maneira semi-aleatória. A seleção buscou eliminar entradas claramente delirantes, a fim de medir a capacidade de sintetização de discurso depressivo em *outputs* bem sucedidos. O número limitado de amostras decorre da inabilidade de realizar os testes em escala, assim como da curadoria feita para eliminar perguntas pobres em conteúdo relevante.

## 3 Resultados

Do ponto de vista quantitativo, o F1-Score foi utilizado como métrica principal para avaliar o impacto dos dados sintéticos no desempenho do classificador de depressão. Os resultados apresentados na Tabela 1 mostram uma melhoria substancial quando modelos ajustados com dados sintéticos são utilizados, especialmente nos estágios intermediários de treinamento. Observa-se que os *checkpoints* 33 e 66 apresentam ganhos expressivos em relação ao conjunto original, indicando que a qualidade dos dados sintéticos gerados nesses estágios contribuiu positivamente para a detecção de discurso depressivo. Por outro lado, o desempenho reduzido no *checkpoint* 100 sugere possível sobreajuste ao domínio sintético ou perda de diversidade na geração, ressaltando a importância de identificar o ponto ideal de convergência para maximizar o benefício do aumento de dados.

Além disso, é importante destacar que, embora os resultados tenham sido consistentes dentro da proposta do estudo, o F1-Score obtido permanece inferior ao reportado em trabalhos de ponta na literatura. Isso se deve em grande parte ao fato de que nosso classificador utiliza exclusivamente transcrições textuais, enquanto abordagens mais avançadas empregam modalidades adicionais — como áudio e vídeo — que capturam prosódia, expressões faciais e outros marcadores comportamentais relevantes para a predição de níveis de depressão. Assim, o uso de uma única modalidade textual naturalmente limita o desempenho máximo possível do sistema, o que reforça o desafio inerente ao cenário estudado.

O escore feito pelo *Chat-GPT 4* leva em consideração o contexto próximo da conversa, estabelecendo relação causal entre a resposta sintética e a pergunta feita, inseridos na dinâmica de uma consulta psiquiátrica/psicológica real. A Tabela 1 apresenta o registro dos resultados dos três modelos avaliados.

Além das métricas quantitativas, realizamos também uma avaliação qualitativa das respostas geradas pelos três modelos, conduzida diretamente pelos desenvolvedores. Nessa análise, verificamos se

Table 1: Desempenho dos modelos após aumento de dados sintéticos

| **Modelo (%)** | **Descrição** | **Depressão F1-Score** | **GPT-Score / 10** |
|---|---|---|---|
| N/A | Conjunto sem dados sintéticos (baseline) | 0.0833 | N/A |
| Checkpoint 33 | Ajuste inicial | 0.4667 | 0.770 |
| Checkpoint 66 | Ajuste intermediário | 0.5238 | 0.775 |
| Checkpoint 100 | Ajuste completo | 0.3636 | 0.795 |

os enunciados sintéticos eram coerentes com as perguntas originais e compatíveis com o nível de discurso depressivo esperado. De modo geral, as respostas apresentaram boa consistência semântica e mantiveram alinhamento com o contexto das interações. Entretanto, observamos alguns comportamentos indesejados, como repetições excessivas, dificuldade do modelo em finalizar a resposta mesmo com o uso do token especial [END] e, ocasionalmente, a produção de trechos em outros idiomas. Curiosamente, quando traduzidos, esses trechos mantinham sentido em relação à pergunta, sugerindo que tal comportamento pode estar associado à configuração de temperatura utilizada durante a geração. Esses achados qualitativos complementam as métricas formais e ajudam a compreender melhor os limites e potenciais do modelo na criação de dados sintéticos úteis.

## 4 Conclusão e Trabalhos Futuros

Este trabalho investigou o uso de um modelo de linguagem de porte moderado (TinyLLaMA, 1.1B parâmetros) como gerador de dados sintéticos para apoiar a detecção automática de discurso depressivo em um cenário de escassez e desbalanceamento de dados clínicos. Os resultados demonstraram que, mesmo com apenas 0,1% dos parâmetros ajustados via LoRA, o modelo foi capaz de produzir amostras sintéticas úteis, especialmente nos estágios iniciais e intermediários do ajuste fino. Os *checkpoints* 33 e 66 apresentaram ganhos substanciais de F1-score em relação ao conjunto original, indicando que dados gerados por um modelo leve podem, de fato, aumentar a robustez de classificadores treinados em bases clínicas reduzidas. A análise qualitativa reforçou essa conclusão ao mostrar que, embora algumas limitações persistam — como repetições, dificuldade no encerramento de respostas e inserção esporádica de trechos em outros idiomas — a maioria das respostas manteve coerência com o contexto das perguntas e compatibilidade com os níveis de discurso depressivo esperados.

No entanto, também foram observadas limitações importantes. O desempenho reduzido no *checkpoint* 100 sugere que estágios mais avançados de treinamento podem levar à perda de diversidade sintética ou ao sobreajuste a padrões específicos do conjunto original. Isso ressalta a necessidade de identificar um ponto de convergência que maximize a utilidade das amostras geradas sem comprometer sua variabilidade linguística. Além disso, embora o uso de uma LLM robusta (*GPT-4*) como avaliadora tenha fornecido indícios valiosos sobre a qualidade sintética, a escala da avaliação permaneceu limitada, e a dependência de um modelo maior pode levantar questões sobre reprodutibilidade em ambientes com restrições computacionais, que são o foco do projeto.

Como trabalhos futuros, diversas direções se mostram promissoras. Primeiramente, é essencial ampliar a avaliação qualitativa com a participação de especialistas em saúde mental, permitindo verificar a adequação clínica das respostas geradas. Em paralelo, explorar diferentes configurações de geração — como temperatura, *top-k*, *top-p* pode reduzir erros recorrentes e aumentar a fluência das amostras. Outra linha relevante envolve comparar outros modelos compactos (500M–3B parâmetros) para investigar como características arquiteturais influenciam a qualidade dos dados sintéticos. Estudos futuros também poderiam avaliar outras formas de aumento de dados que complementem o aumento dos textos, como abordagens multimodais de áudio e vídeo, que também são muito utilizadas em modelos preditores de níveis de depressão.

Em síntese, este estudo mostra que modelos de linguagem leves representam uma alternativa promissora para expansão de bases clínicas sensíveis, especialmente em contextos nos quais privacidade, custo computacional e dificuldade de coleta são fatores críticos. O uso controlado de dados sintéticos gerados por tais modelos pode, portanto, contribuir para o avanço de ferramentas assistivas em saúde mental de forma segura, eficiente e acessível.

## References


[1] G. Ansari et al. Data augmentation for mental health classification on social media. In *Proceedings of ICON 2021*, 2021.

[2] J. Aranha and Others. A comprehensive review of datasets for clinical mental health. *arXiv preprint*, 2025.

[3] Chbmleao. Finetune-LLM-Depression-Data: Synthetic data generation for depression detection via Parameter-Efficient Fine-Tuning of Large Language Model. `https://github.com/Chbmleao/Finetune-LLM-Depression-Data`, 2025. Versão consultada em 25-nov-2025.

[4] B. Ding. Data augmentation using large language models. *arXiv preprint*, 2024.

[5] USC Institute for Creative Technologies. Daic-woz dataset for depression and ptsd assessment. `https://dcapswoz.ict.usc.edu/`, 2014. Dataset documentation.

[6] Centers for Disease Control and Prevention. Depression prevalence databrief (2021–2023). `https://www.cdc.gov/nchs/products/databriefs/db527.htm`, 2023. Accessed: 2025-01-15.

[7] L. Gómez-Zaragozá et al. Speech and text foundation models for depression detection. In *Proceedings of Interspeech 2025*, 2025. To appear.

[8] I. Lorge et al. Detecting the clinical features of difficult-to-treat depression: synthetic data to alleviate data scarcity. *Journal on Biomedical Informatics*, 2025. Forthcoming.

[9] D. E. Merzougui et al. Evaluating large language models for depression symptom classification. In *AIME 2025 – Artificial Intelligence in Medicine*, 2025. To appear.

[10] National Institute of Mental Health. Major depression — statistics. `https://www.nimh.nih.gov/health/statistics/major-depression`, 2021. Accessed: 2025-01-15.

[11] World Health Organization. Depressive disorder (depression) — fact sheet. `https://www.who.int/news-room/fact-sheets/detail/depression`, 2025. Accessed: 2025-01-15.

[12] J. Ye et al. Llm-da: Data augmentation via large language models. *arXiv preprint*, 2024.