

**Tech Challenge 3**

**IA para Devs**

**Daniela Cruz de Malta - RM353365**

**Gabriela Maciel Godoi - RM355125**

**Lucas Sutelo - RM353721**

**Sumário**

Caso de Estudo [3](#_heading=h.yws6ghm3o5ej)

Metodologia [3](#_heading=h.yws6ghm3o5ej)

[1. Instalação de Biblioteca Necessárias 3](#_heading=h.yws6ghm3o5ej)

2. Configuração do Ambiente3

3. Escolha do Dataset3

4. Chamada do Foundation Model 3

5[.Processo de fine-tuning 3](#_heading=h.yws6ghm3o5ej)

6. Documentação e Salvamento do modelo  [3](#_heading=h.yws6ghm3o5ej)

Resultado [3](#_heading=h.yws6ghm3o5ej)

Conclusão [3](#_heading=h.yws6ghm3o5ej)

Anexo [3](#_heading=h.yws6ghm3o5ej)

# 

O objetivo deste Tech Challenge é realizar o fine-tuning de um foundation model utilizando a base de dados “The AmazonTitles-1.3MM” para receber perguntas dentro do contexto da base pelo usuário dentro do Prompt e a resposta conter os aprendizados do fine-tuning.

Este documento detalha o processo de seleção e preparação da base de dados e descreve o processo de fine-tuning do modelo, com os parâmetros e ajustes utilizados.

O código fonte do processo de fine-tuning está no anexo do documento.

# Metodologia

## 1. [Instalação de Biblioteca Necessárias](#_heading=h.yws6ghm3o5ej)

* As bibliotecas *accelerate*, *peft*, *bitsandbytes*, *transformers* e *trl* são instaladas para fornecer as ferramentas necessárias para realizar o *fine-tuning* e a quantização de modelos.

## 2. Configuração do Ambiente

## O código importa PyTorch e outros módulos da biblioteca Hugging Face para manipulação de modelos e datasets.2. Processo de fine-tuning

* As classes **AutoModelForCausalLM**, **AutoTokenizer**, e **SFTTrainer** são utilizadas para carregar o modelo base e realizar o *fine-tuning*.

3. Escolha do Dataset

* O dataset utilizado para o treinamento é o **AmazonTitles-1.3MM**, carregado a partir do arquivo **trn.json**, contendo títulos e descrições de produtos.
* As colunas "title" e "content" são preparadas para formar os *prompts* de entrada no modelo.

4. Chamada do Foundation Model

* O modelo base escolhido é o **Llama-2-7b-chat-hf**, carregado da plataforma **Hugging Face**.
* Um teste inicial é executado para medir a performance do modelo antes do *fine-tuning*.

5. [Processo de fine-tuning](#_heading=h.yws6ghm3o5ej)

* O modelo é configurado para executar o *fine-tuning* utilizando a quantização em 4 bits através da configuração **BitsAndBytesConfig**.
* O processo de ajuste fino utiliza a técnica **LoRA** (*Low-Rank Adaptation*) para treinar o modelo em menor escala, economizando recursos computacionais.
* Os parâmetros de treinamento incluem 15 épocas, tamanho de lote de 4, taxa de aprendizado de 0.001 e otimizador *Paged AdamW*.

6.

# Resultados

## 1. Antes do fine-tuning

lalala

## 2. Após o fine tuning

lalala

# Conclusão

O código cumpre o objetivo do **Tech Challenge** de realizar o *fine-tuning* de um modelo de linguagem grande utilizando o dataset **AmazonTitles-1.3MM**. O fluxo de trabalho documentado garante que o modelo esteja preparado para responder perguntas baseadas em descrições de produtos, oferecendo uma base sólida para aplicações em processamento de linguagem natural no contexto de comércio eletrônico.

# Anexo

Código fonte.