

# **Projeto Inteligência Artificial**

**Universidade Fernando Pessoa  
2024/2025**



**Chat Bot**

**Bruno Silva nº42824  
Lucas Gonçalves nº42380  
Manuel Gonçalves nº43087**

## Índice

Resumo .....	3
Introdução .....	4
1. Objetivos do trabalho .....	4
2. Motivação .....	5
Descrição do Problema .....	6
Estado da Arte .....	7
Descrição do trabalho realizado .....	8
Análise de Resultados .....	9
1. Treino do modelo (LORA) .....	9
2. RAG .....	10
3. RAG e LORA .....	11
4. Histórico .....	12
Conclusão .....	13

## Resumo

Este projeto tem como objetivo o desenvolvimento de um chatbot inteligente em inglês baseado em modelos de linguagem de grande escala (LLMs), utilizando o Gemma 2B, com o propósito de divulgar a Universidade Fernando Pessoa (UFP) e os seus cursos de Engenharia Informática (INF) e Mestrado em Informática (MINF).

Foram aplicadas técnicas modernas de afinação, como LoRA e few-shot learning, bem como Retrieval-Augmented Generation (RAG), para melhorar a precisão das respostas.

Os resultados demonstraram respostas relevantes abrindo caminho para futuras aplicações de apoio a alunos e promoção institucional da UFP.

## Introdução

Este projeto explora a utilização de modelos de linguagem de grande escala (LLMs), em particular o Gemma da Google, para desenvolver um chatbot destinado a promover a Universidade Fernando Pessoa (UFP) e os seus cursos de Engenharia Informática (INF) e Mestrado em Informática (MINF). A abordagem adotada combina técnicas modernas, como LoRA para um ajuste eficiente, few-shot learning para adaptação com poucos exemplos e RAG (Retrieval-Augmented Generation) para melhorar a precisão das respostas com base em dados externos. O sistema será testado num ambiente local ou em nuvem.

### 1. Objetivos do trabalho

Os principais objetivos deste projeto são:

- Desenvolver um chatbot informativo sobre a UFP e os seus cursos de INF e MINF, recorrendo ao modelo Gemma 2B ou 7B;
- Implementar técnicas de fine-tuning com LoRA e few-shot learning para adaptar o modelo ao domínio universitário.
- A integração de Retrieval-Augmented Generation (RAG) visa reforçar a fiabilidade e a contextualização das respostas.
- Testar e validar o sistema num ambiente local/cloud.

## **2. Motivação**

Com o avanço das tecnologias de inteligência artificial, os modelos de linguagem de grande escala (LLMs) transformaram-se em ferramentas poderosas na construção de sistemas de conversação inteligentes. A aplicação destes modelos no contexto académico proporciona uma oportunidade única de promover instituições de ensino superior de maneira inovadora, acessível e personalizada. A possibilidade de criar um chatbot com compreensão de linguagem natural, adaptado ao domínio da Universidade Fernando Pessoa (UFP), representa um passo significativo na modernização da comunicação institucional.

## Descrição do Problema

O desafio consiste em desenvolver um sistema de conversação personalizado, centrado na divulgação institucional e académica da UFP. O sistema deve:

- Compreender e responder a perguntas sobre a UFP, o INF e o MINF;
- Deve ser personalizável e escalável para contextos educacionais.

Este problema envolve a seleção de dados relevantes, a adaptação de um modelo de linguagem de grande escala ao domínio específico da UFP.

## Estado da Arte

Nos últimos anos, os LLMs, como o GPT, o BERT, o T5, o Gemini e agora o Gemma, têm revolucionado a interação entre seres humanos e máquinas. O modelo Gemma, desenvolvido pela Google DeepMind, oferece variantes compactas (2B e 7B) e é compatível com ferramentas como Keras, TensorFlow e Hugging Face, sendo ideal para projetos acadêmicos e aplicativos com limitações de recursos computacionais.

Entre as técnicas utilizadas, destacam-se:

- LoRA: permite o ajuste fino eficiente de LLMs com baixo custo computacional;
- Few-shot Learning: reduz a necessidade de grandes quantidades de dados anotados;
- RAG (Retrieval-Augmented Generation): melhora a precisão das respostas recorrendo a fontes externas de conhecimento dinâmico.

Adicionalmente, a Google Mediapipe fornece uma estrutura de inferência para LLMs em dispositivos móveis, essencial para a fase de implementação do Android.

## Descrição do trabalho realizado

O trabalho foi dividido em várias fases:

- Preparação dos dados: foram recolhidos documentos e informações relevantes da UFP e dos cursos INF/MINF, estruturando-os em formatos apropriados (por exemplo, JSONL).
- Configuração do modelo Gemma 2B:
  - Utilização de LoRA para afinação do modelo com base no domínio universitário.
  - Aplicação de few-shot learning com exemplos orientados para perguntas frequentes.
- Integração com RAG:
  - Implementação de um sistema de recuperação com base em embeddings semânticos (usando FAISS).
  - Geração de respostas fundamentadas com base no conteúdo recuperado.
- Testes locais (PC/Cloud):
  - Validação do desempenho do modelo em testes de perguntas e respostas.



## Análise de Resultados

### 1. Treino do modelo (LORA)

```

**** Running training ****
Num examples = 105
Num Epochs = 3
Instantaneous batch size per device = 1
Total train batch size (w. parallel, distributed & accumulation) = 4
Gradient Accumulation steps = 4
Total optimization steps = 81
Number of trainable parameters = 20,766,720
0%|
/Users/manuelgoncalves/.virtualenvs/projectAI/lib/python3.13/site-packages/torch/utils/data/dataloader.py:683: UserWarning: 'pin_memory' argument
e but not supported on MPS now, then device pinned memory won't be used.
  warnings.warn(warn_msg)
It is strongly recommended to train Gemma2 models with the 'eager' attention implementation instead of 'sdpa'. Use 'eager' with 'AutoModelForCausalLM'
ained('<path-to-checkpoint>', attn_implementation='eager')'.
{'loss': 1.9192, 'grad_norm': 2.8213202953338623, 'learning_rate': 0.0001970941817426052, 'epoch': 0.30}
20%|

```

A print mostra o processo de aprendizagem do modelo (estando, no momento em que a print foi tirada, nos 20%).

```

1. RAG
2. LORA
3. RAG + LORA
Choose an option (1/2/3): 2

Your question: List the courses of the first year of the Bachelor's Degree in Computer Engineering at UFP?
Using LoRA for response generation...

Answer: The courses are: 1st Semester: Applied Statistics (5 ECTS), English I (3 ECTS), Introduction to Algorithmics and Programming (6 ECTS), Mathematics I (4 ECTS), Physics (4 ECTS); 2nd Semester: Systems Analysis (6 ECTS), Applied Electronics (7 ECTS), Mathematics II (7 ECTS), Option 1 (6 ECTS). Total: 60 ECTS.

Your question: How much ECTS does Applied Statistics from computer engineering UFP degree have?
Using LoRA for response generation...

Answer: Applied Statistics in the Bachelor's Degree in Computer Engineering at UFP has 6 ECTS.

Your question:

```

A print mostra o modelo a funcionar com Lora (+ historico).

## 2. RAG

```
loading checkpoint shards: 100% | 1/2 100.0/500.0M
Welcome to the UFP Chatbot! Ask about UFP courses or type 'quit' to exit.

Your question: List the courses of the first year of the Bachelor's Degree in Computer Engineering at UFP?

Answer: The courses of the first year of the Bachelor's Degree in Computer Engineering at UFP with their ECTS credits are:
1st Semester:
- Applied Statistics: 5 ECTS
- Communication Grammar: 3 ECTS
- English: 3 ECTS
- Introduction to Algorithmics and Programming: 6 ECTS
- Mathematics I: 7 ECTS
- Information Systems: 6 ECTS
2nd Semester:
- Systems Analysis: 6 ECTS
- Applied Electronics: 7 ECTS
- Physics: 4 ECTS
- Mathematics II: 7 ECTS
- Option 1: 6 ECTS
Total: 60 ECTS

Here's an elaboration:

**1st Semester:**

Your question:

Answer: The course Multimedia Communication in the Bachelor's Degree in Communication Sciences at UFP has 6 ECTS.

**Explanation:**

The answer provides the requested information directly. It is concise and accurate based on the provided context.

Your question: █
```

A print mostra que a resposta gerada apresenta uma estrutura clara, com segmentação por semestre e o número de créditos ECTS corretos. O sistema também fornece respostas diretas. (RAG com histórico).

### 3. RAG e LORA

```
1. RAG
2. LORA
3. RAG + LORA
Choose an option (1/2/3): 3

Your question: List the courses of the first year of the Bachelor's Degree in Computer Engineering at UFP?
Using RAG to retrieve relevant documents and LoRA for response generation...

Answer: The courses of the first year of the Bachelor's Degree in Computer Engineering at UFP are: 1st Semester: Applied Statistics (5 ECTS), Communication Grammar (3 ECTS), English (3 ECTS), Introduction to Algorithmics and Programming (6 ECTS), Mathematics I (7 ECTS), Information Systems (6 ECTS); 2nd Semester: Systems Analysis (6 ECTS), Applied Electronics (7 ECTS), Physics (4 ECTS), Mathematics II (7 ECTS), Option 1 (6 ECTS). Total: 60 ECTS.

Your question: How much ECTS does Applied Statistics from computer engineering UFP degree have?
Using RAG to retrieve relevant documents and LoRA for response generation...

Answer: Applied Statistics in the Bachelor's Degree in Computer Engineering at UFP has 5 ECTS.
```

A print mostra o modelo a funcionar com RAG e LORA (juntamente com o histórico).

#### 4. Histórico

Your question: How mutch ECTS does Applied Statistics from computer engineering UFP degree have?

Answer: Applied Statistics has 5 ECTS credits in the Bachelor's Degree in Computer Engineering at UFP.

Explanation:

The answer is extracted from the context provided. It is a direct answer to the question.

Your question: ■

Mostra a capacidade do sistema de responder de forma consistente, indicando também que há coerência na resposta, mesmo em caso de reformulação da pergunta.

## Conclusão

Este projeto demonstrou a viabilidade de utilizar LLMs, como o Gemma 2B, para construir um chatbot institucional. Com as ferramentas modernas de afinação e recuperação, foi possível obter um desempenho robusto em tarefas de conversação orientadas para o domínio educacional.

Para o futuro, propõe-se o seguinte:

- Conclusão da integração com dispositivos Android via Mediapipe;
- Expansão do conhecimento do chatbot para incluir todos os cursos e serviços da UFP;