

LAW TALK

Data Mining



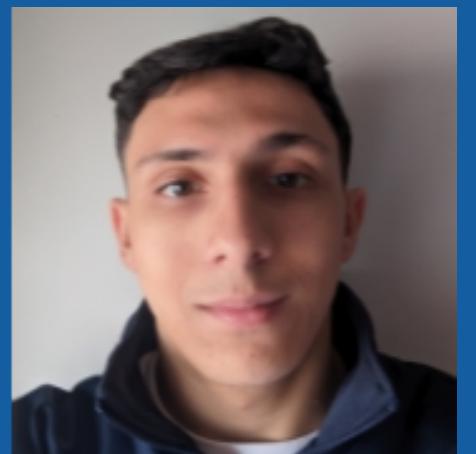
Afonso Bessa

PG53597



Eduardo Henriques

pg54780



Francisco Claudino Mariana Marques

PG50380

PG50633



24/05/2024

INDEX

01

Introduction

02

Architecture

03

Data Sources

04

Data Extraction
and Database

05

Graphical Interface

06

Result Analysis

INTRODUCTION

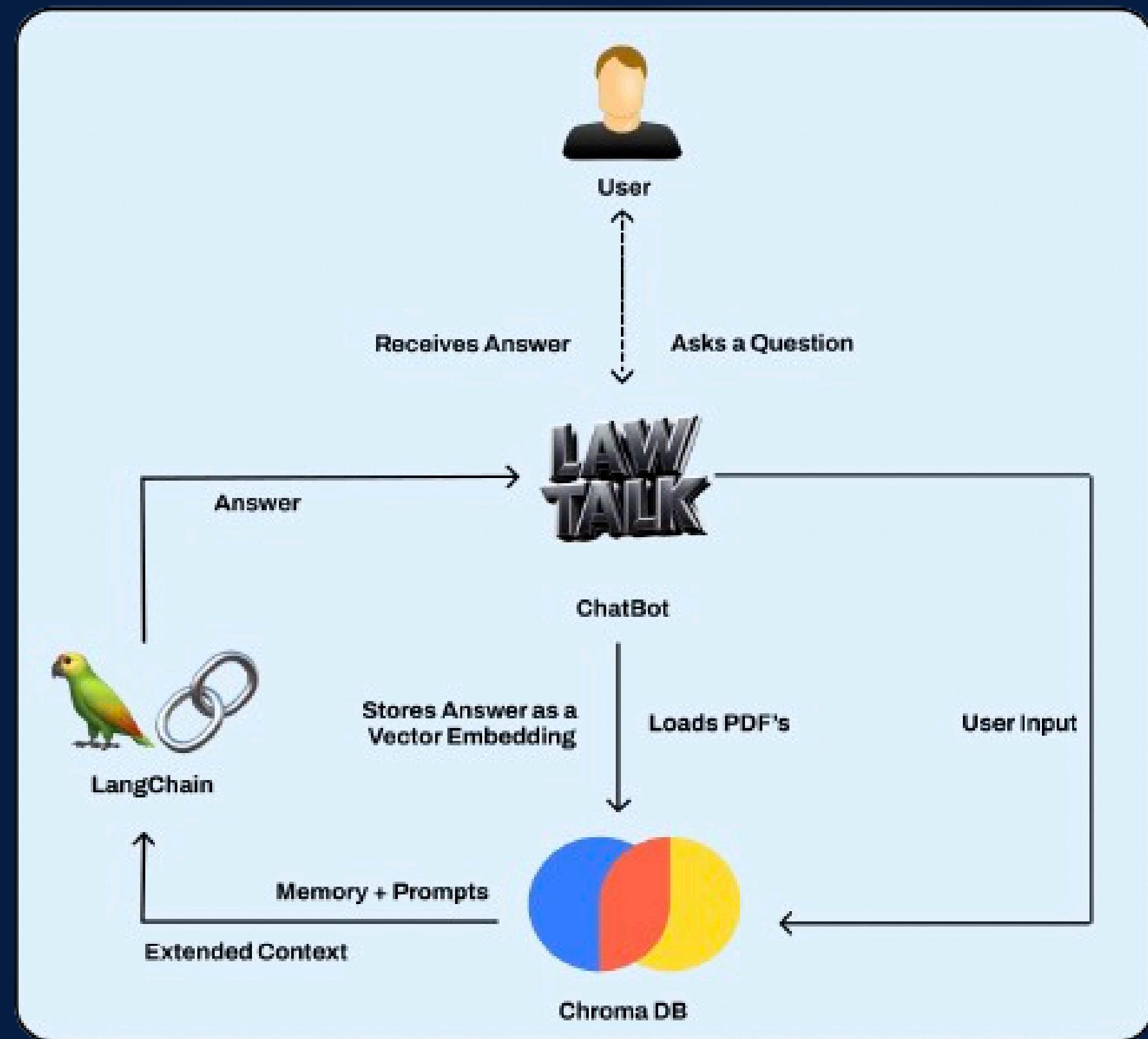
CONTEXTUALIZATION

- O1** Chatbot capable of solving practical cases corresponding to crimes, providing information about the legal consequences to be applied to the offender.
- O2** Chatbot capable of citing the articles on which it relied to solve the practical cases.

OBJECTIVES

- O1** Practical ChatBot that answers in Natural Language
- O2** Customization in terms of LLM model options and embeddings.

ARCHITECTURE



DEVELOPMENT PLAN RESOURCES

01 Language - Python

02 Framework - Langchain

01 Vector Database - ChromaDB

02 Evaluators - String Evaluators

03 Large Language Model - Ollama Family

04 User Interface - Streamlit



Ollama

DATA SOURCES

PDF'S regarding the:

- 01 Portuguese Penal Code
- 02 Portuguese Processual Penal Code
- 03 Practical Cases and respective Answers
- 04 Notes taken from Portuguese Law Classes



DATA EXTRACTION

EXTRACTION TOOLS

Py2PDF

- Detailed Customization
- Limited features for extracting data
- Only suitable to extract data from images

PDFPlumber

- Handles Complex Layouts
- High Accuracy in text and table data extraction
- Can handle both Scanned and Digital PDF'S

Tesseract

- Requires image pre-processing
- More Suitable to image based PDF's
- Poor performance on PDF's with Complex Layouts

DATA PROCESSING

01 Data Processing of TXT Files

- 01 Clean the Text
- 02 Character Normalization
- 03 Tokenization
- 04 Indentation



DATABASE

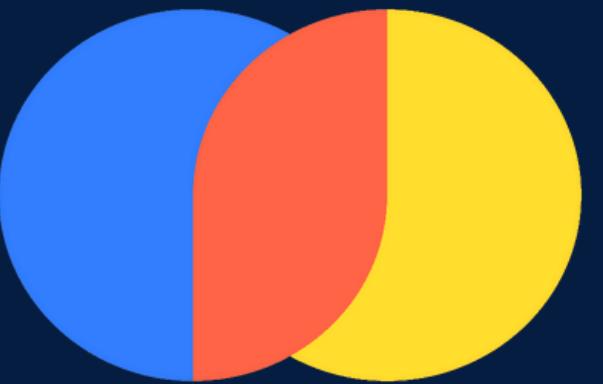
CHROMADB

Storing PDF's and Vector Embeddings

01 Stores the PDF's as soon as it initializes

02 In each question, the Retriever provides Documents for extra context along with the User Input

03 In each response given by the LLM, it stores the Answer in a Vector Embedding



ChromaDB

EXAMPLE

Question:

As finalidades do processo penal podem ser entendidas em que sentido?



Answer:

*As finalidades do processo penal podem ser entendidas em dois sentidos: o **sentido axiológico** e o **sentido sociológico**.*



GRAPHICAL INTERFACE

Streamlit

The screenshot shows a Streamlit application window titled "LAW TALK". On the left, there's a sidebar with "Settings" and a checked "Reload Embeddings" option. Below it is an "LLM Model Name" dropdown menu with "llama2" selected, and other options like "mistral", "llama2", and "zephyr". The main area features a dark background with a gavel icon and text about Portuguese Penal Code articles. A button at the bottom says "Envia Mensagem ao LawTalk >".

Deploy :

LAW TALK

Settings

Reload Embeddings

LLM Model Name

llama2

mistral

llama2

zephyr

Em que casos se podem aplicar os artigos do Livro 1, Título 4 do Código Penal Português ?

Com base na pesquisa fornecida, os artigos do Livro 1, Título 4 do Código Penal Português podem ser aplicados nos seguintes casos:

1. Infrações puníveis por lei: A vítima ou seus sucessores têm o direito de apresentar uma queixa privada por crimes puníveis por lei, como assassinato, roubo, danos à propriedade, difamação e outros.
2. Difamação: Indivíduos que foram difamados também podem apresentar uma reclamação privada nos termos do Artigo 149, buscando indenização por danos causados à sua reputação. Portanto, esses artigos podem ser aplicados em casos de assassinato, roubo, danos à propriedade, difamação e outros crimes. Esta é uma questão importante para garantir que as vítimas e as partes afetadas tenham um meio de buscar justiça nos casos em que os promotores públicos podem não priorizá-los ou podem ser incapazes de fazê-lo devido a limitações de recursos ou outros motivos.

Envia Mensagem ao LawTalk >

BENCHMARKING

Accuracy and Precision Measurements

O1 Use of an Evaluator from LangChain

O2 Creation of a Test Dataset with question and expected answer

O3 Creation of Quantity and Quality Criteria

- Scores between 1 and 10
- With each score comes a quality evaluation
- Quality evaluation based on accuracy and precision



LangChain

RESULT ANALYSIS

Mistral

- Scores between 1 and 10
- Extremely inconsistent with the answers
- Susceptible to hallucinations
- Answer Time : 25 to 35 secs

Llama2

- Scores between 7 and 10
- Very consistent with the answers
- Not susceptible to hallucinations
- Answer Time : 30 to 40 secs

Zephyr

- Scores between 5 and 10
- Mostly consistent with the answers
- Not susceptible to hallucinations
- Answer Time : 35 to 45 secs



Ollama

LAW TALK

Data Mining



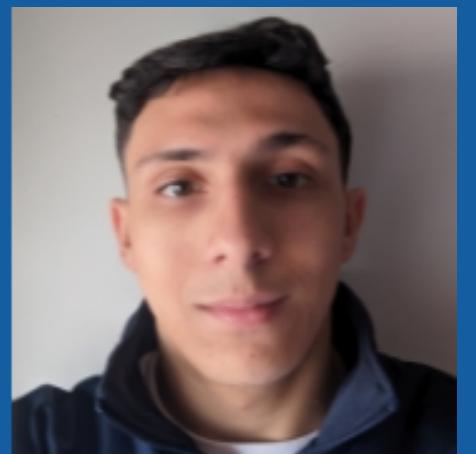
Afonso Bessa

PG53597



Eduardo Henriques

pg54780



Francisco Claudino Mariana Marques

PG50380

PG50633



24/05/2024