AI-Powered Real Estate Assistant

Project Overview

We aim to develop a custom AI model tailored to the real estate industry in Germany. The assistant should process unstructured datasets, legal documents, and visual data, learn domain-specific knowledge, and provide professional, interactive responses in German. The model will integrate advanced data extraction, natural language understanding, and image recognition to deliver comprehensive assistance.

Key Objectives

1. Develop an AI model specializing in German real estate knowledge.
2. Implement advanced data extraction from PDFs and text documents.
3. Incorporate text processing, natural language understanding (NLU), and image recognition for extracting relevant visual data.
4. Integrate extracted information into a cohesive, interactive AI model capable of answering queries, summarizing information, and analyzing real estate-specific scenarios.
5. Ensure efficient fine-tuning using methods like LoRA/QLoRA for computational efficiency.

Technical Requirements

1. Language and Base Model

- **Language**: German
- **Base Model**: Use a German-specific NLP model:
  - Suggested Models:
    - **"bert-base-german-cased"**

        - **"gottbert-base"**
        - **"xlm-roberta-base"** (multilingual with strong German capabilities)
    - Task Types:
        - Question Answering
        - Document Summarization
        - Information Extraction and Classification

---

2. Data Processing and Advanced Features

2.1. Advanced Data Extraction

- Extract structured and unstructured data from PDFs, such as:
    - Legal documents (e.g., tenancy laws, purchase agreements).
    - Market reports and property listings.
- Tools and Libraries:
    - **PyPDF2** or **PDFPlumber** for text extraction.
    - Optical Character Recognition (OCR) for scanned PDFs (e.g., using **Tesseract** or **Adobe PDF Services API**).

2.2. Text Processing

- Use advanced NLP techniques for:
    - Semantic understanding of legal and real estate documents.
    - Summarizing lengthy documents into concise, user-friendly insights.

2.4. Integration

- Consolidate extracted textual, numerical, and visual data into a unified AI model for seamless querying and interaction.

---

3. Fine-Tuning Approach

- **Fine-Tuning Method**:
    - Use **LoRA (Low-Rank Adaptation)** or **QLoRA (Quantized LoRA)** for parameter-efficient fine-tuning.

- Optimize for tasks like:
  - Question Answering
  - Document Summarization
  - Text-Image Integration

---

4. Dataset Preparation

4.1. Sources

- **File Types**: PDFs, images (e.g., JPEG, PNG), text files, and spreadsheets.

4.2. Preprocessing

- Text Data:
  - Tokenization, cleaning, and annotation for domain-specific terms.
  - Format datasets for tasks like question answering (e.g., SQuAD format).

---

5. Deliverables

- **Final AI Model**:
  - Fine-tuned German BERT-based model, integrated with text and image understanding.
  - Optimized for LoRA/QLoRA to reduce computational overhead.
- **Interactive Access**:
  - REST API or lightweight UI (e.g., Streamlit, Gradio).
  - Ability to handle multimodal queries combining text and image inputs.
- **Downloadable Package**:
  - Model weights, tokenizer, and configuration files.
- **Documentation**:
  - Usage guide for the AI model and API.
  - Instructions for further fine-tuning or retraining.
  - Detailed explanation of preprocessing pipelines for PDFs, text, and images.

---

6. Functional Requirements

- **Supported Queries**:
  - Text-Based:
    - "Was sind die Schritte für den Erwerb eines Hauses in München?"
    - "Erklären Sie die aktuellen Trends im deutschen Immobilienmarkt."
- **Performance**:
  - Provide accurate, professional responses in German.
  - Summarize lengthy documents into concise answers.
  - Handle both text and image inputs seamlessly.

---

7. Deployment Requirements

- **Environment**:
  - Deployable locally (e.g., MacBook Pro with M1 Pro) or on cloud platforms (AWS, Azure, Google Cloud).
- **Integration**:
  - REST API for system integration.
  - Docker containerization for portability.
- **Scalability**:
  - Design for scalability to accommodate new datasets and features.

---

8. Maintenance and Scalability

- **Continuous Learning**:
  - Periodic retraining with updated data (e.g., new laws, market trends).
- **Modularity**:
  - Ensure modular pipelines for text and image processing for easy future upgrades.