Smart Document.ai

FBRINSIGHT CHATBOT:

Smart Information Retrival System using LLM and RAG

PRESENTED BY

Jamal Ahmed Khan &
Ahmed Raza

OVERVIEW

01	02	03	04
Problem Statement	Existing Solutions	Drawbacks	Proposed Solution
05	06	07	08
Features	Workflow	Progress	Resources Used

PROBLEM STATEMENT

The Federal Board of Revenue (FBR) and other tax-related organizations in Pakistan face significant challenges in efficiently retrieving relevant tax documents and information. Current manual search processes are time-consuming and error-prone, impacting productivity and taxpayer satisfaction. There is a pressing need for an automated, intelligent system to streamline information retrieval and ensure quick access to accurate tax information.

EXISTING SOLUTIONS

- ChatPDF.com
- Sider.ai
- ChatWithPDF.ai
- AskYourPDF.com
- TextCortex
- Hipdf.com

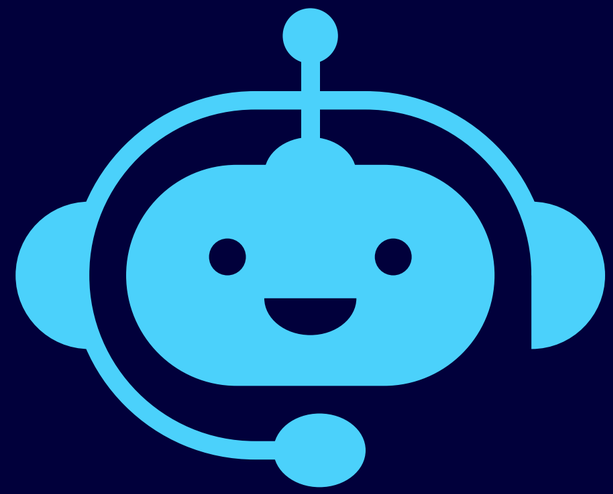
DRAWBACKS

- Limited to pre-defined knowledge base
- Not suitable for complex queries
- Not suitable for confidential documents
- Non Flexible
- Limited Features

PROPOSED SOLUTION

FBRInsight Chatbot:

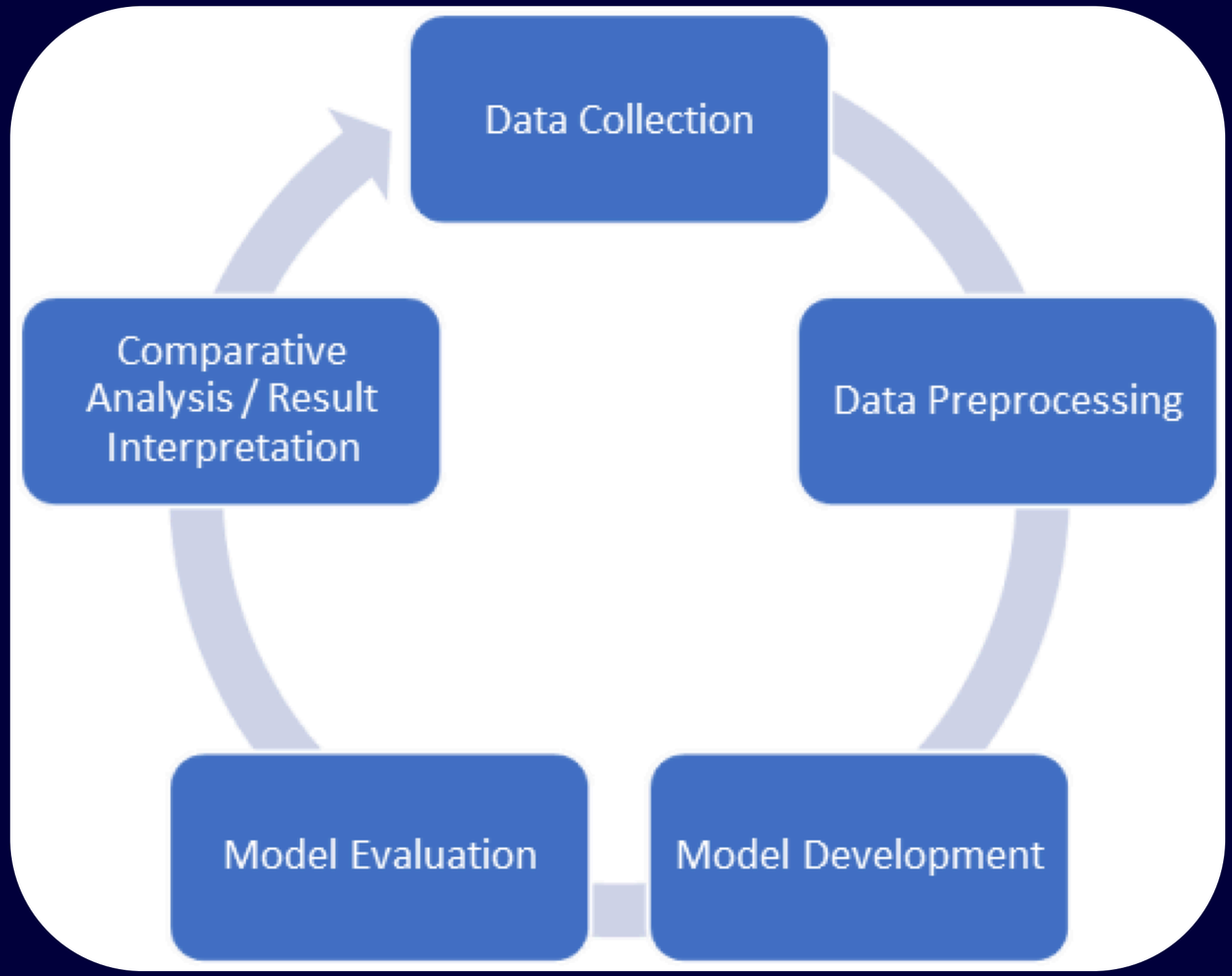
- Streamlines tax document retrieval.
- Utilizes Llama 2 and RAG for accurate responses.
- Pre-processes documents with microsoft/Phi-3-mini-128k-instruct model.
- Employs FAISS for efficient similarity search.
- Built with Streamlit for user interaction.
- Reduces search time, improves accuracy, and boosts productivity.



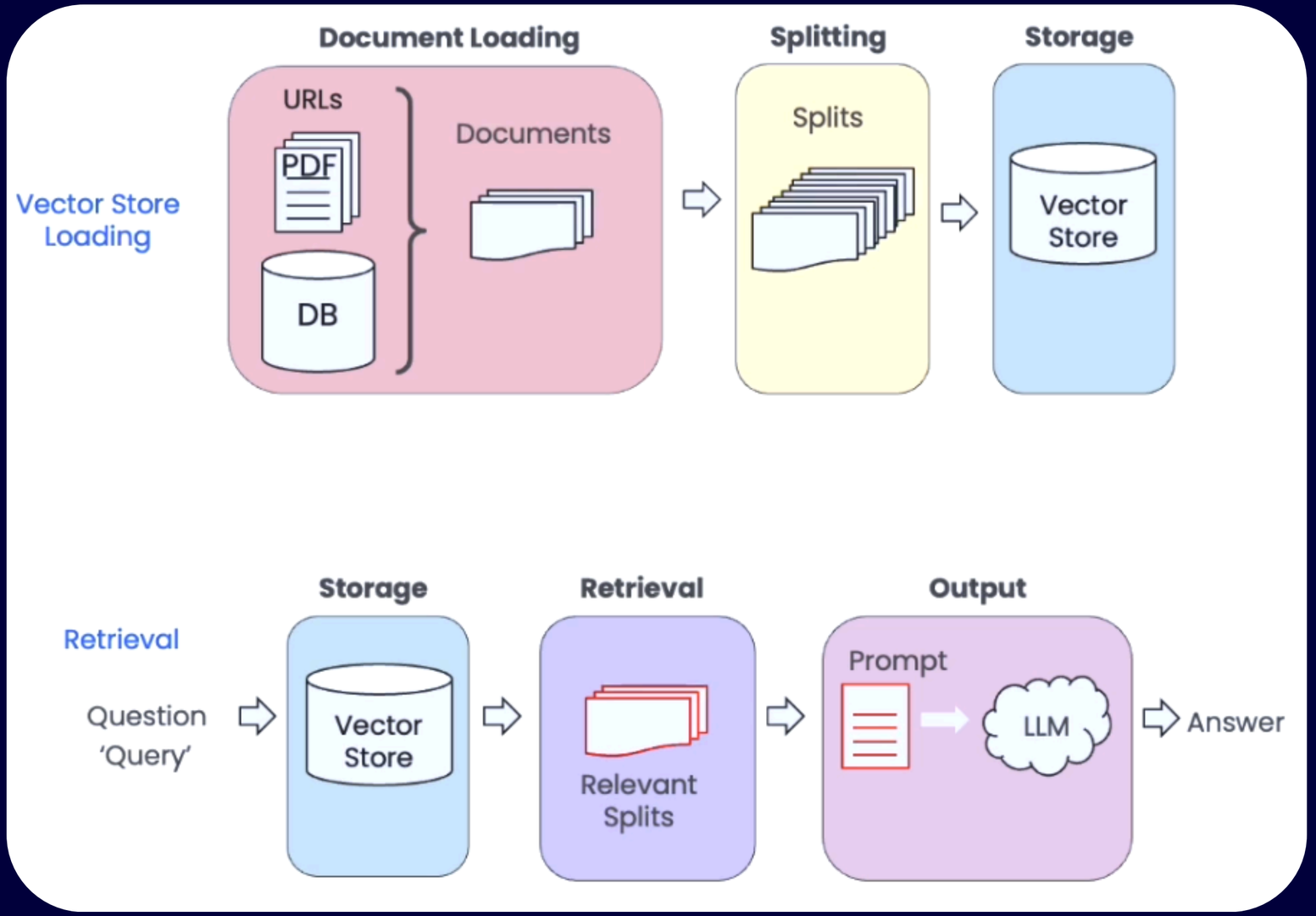
FEATURES

- LLM and RAG
- Advance Technology
- Highly Accurate
- Relevancy Through Information Retrieval
- Confidentiality of Organizational Data
- Domain specific Customization
- On Premises Deployment

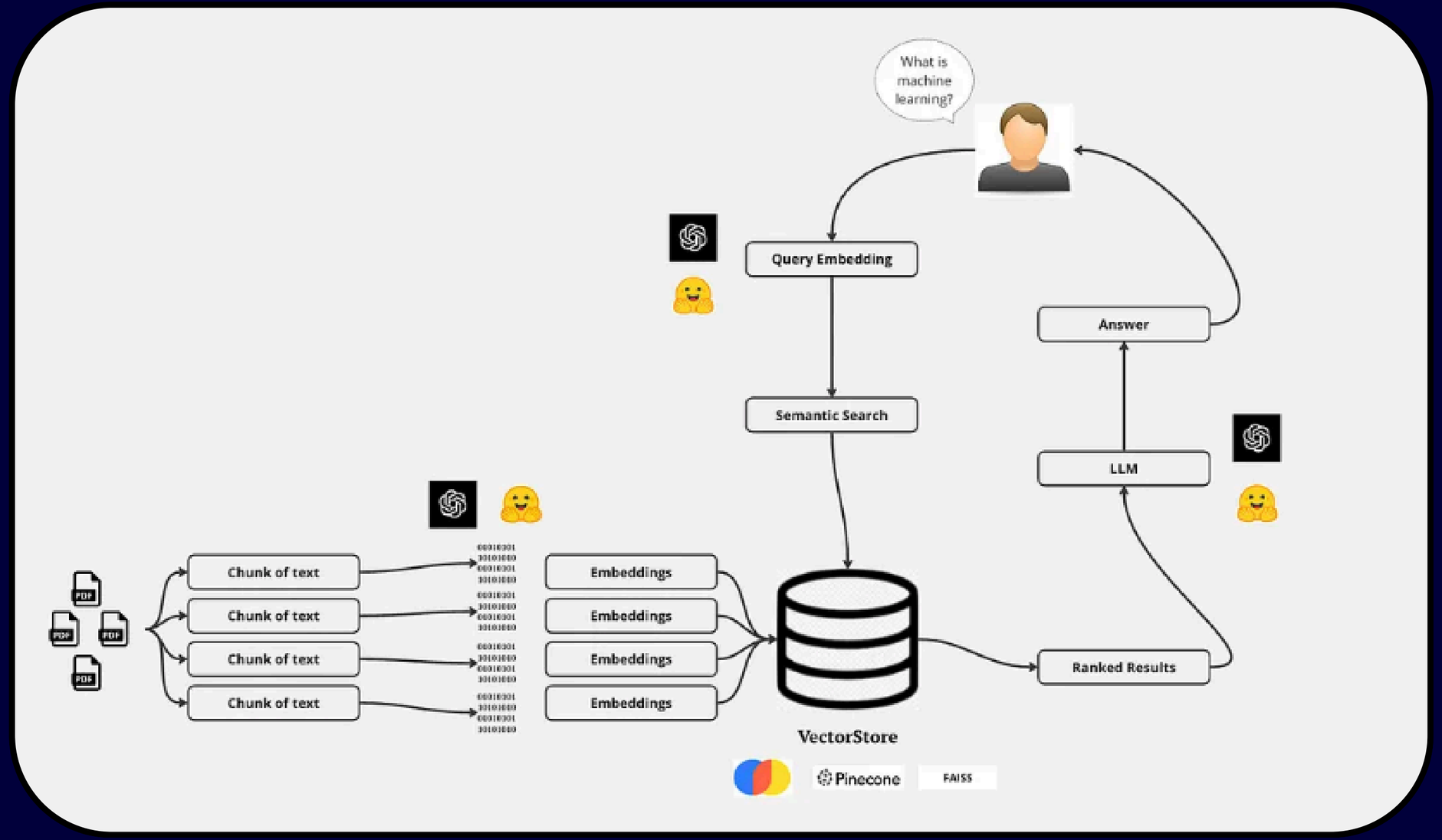
WORKFLOW


Data Collection
Data Preprocessing
Model Development
Model Evaluation
Comparative Analysis / Result Interpretation

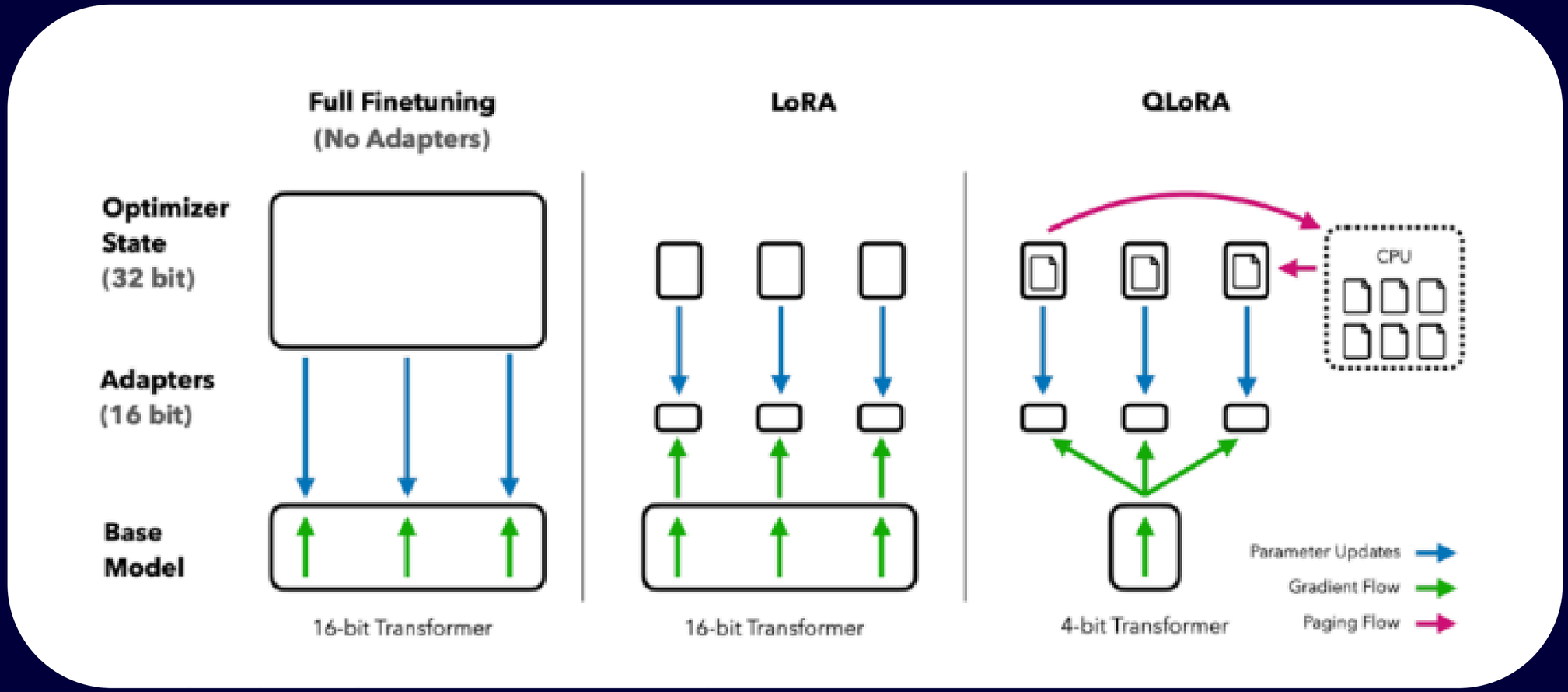
METHODOLOGY



PIPELINE


What is machine learning?
Query Embedding
Semantic Search
Answer
LLM
Chunk of text
Chunk of text
Chunk of text
Chunk of text
Embeddings
Embeddings
Embeddings
Embeddings
VectorStore
Ranked Results
Pinecone
FAISS

LORA CONFIGURATION

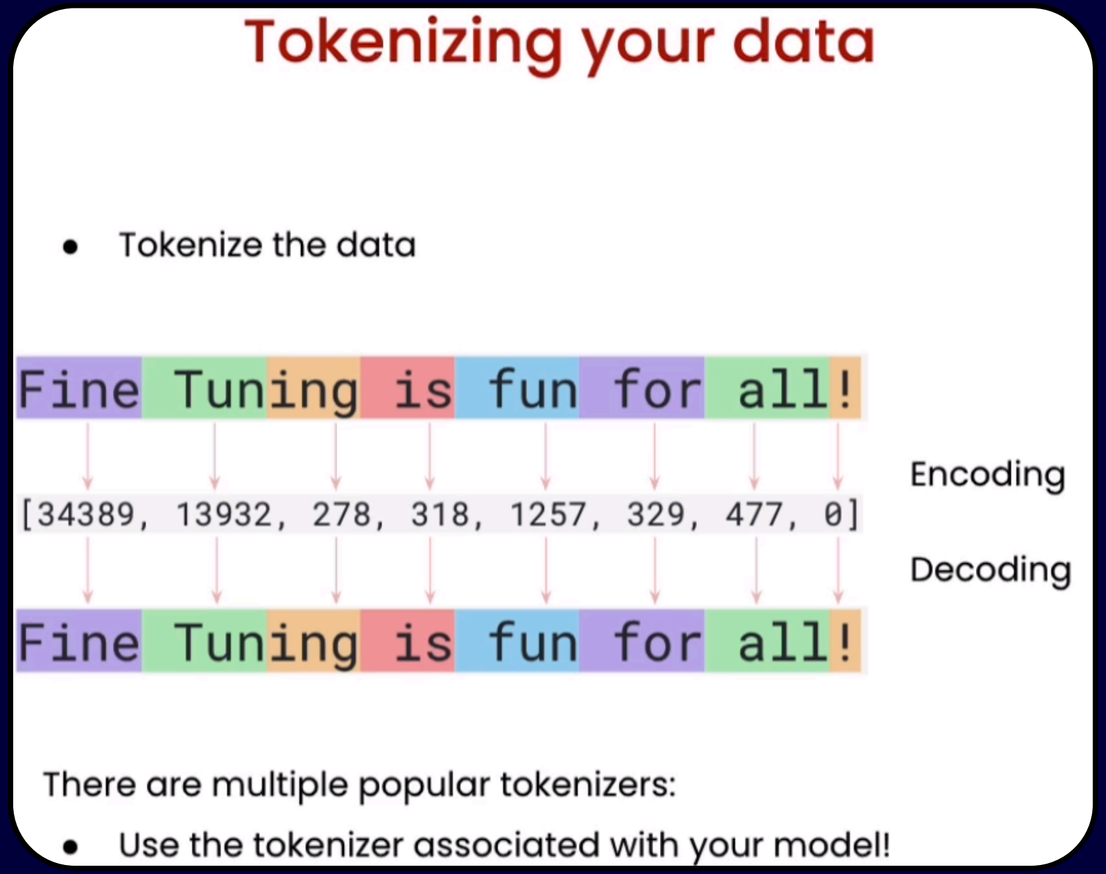


Data Preperation

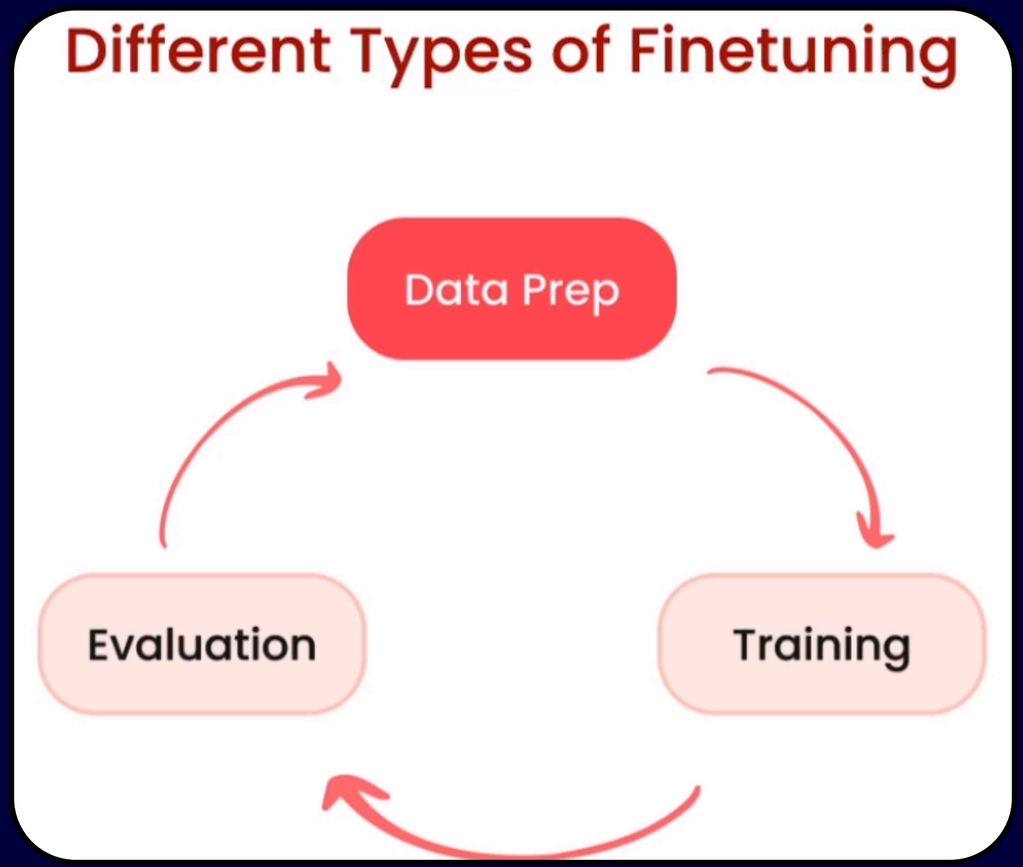
Steps to prepare your data

1. Collect instruction-response pairs
2. Concatenate pairs
   (add prompt template, if applicable)
3. Tokenize: Pad, Truncate
4. Split into train/test

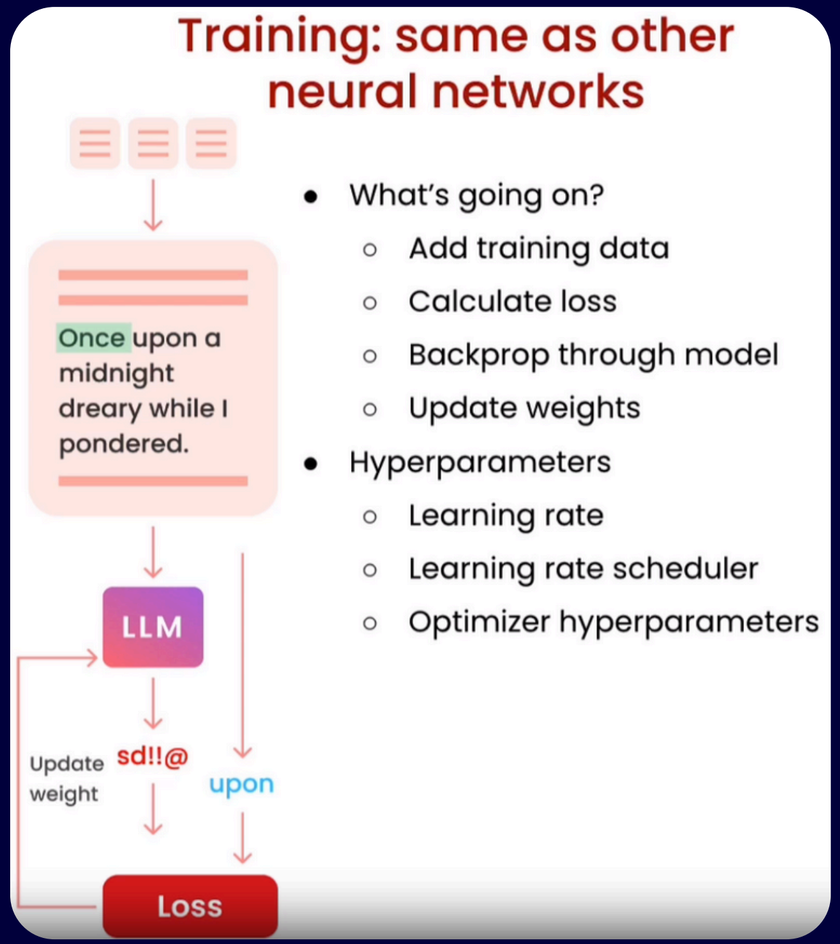
Tokenize the Data



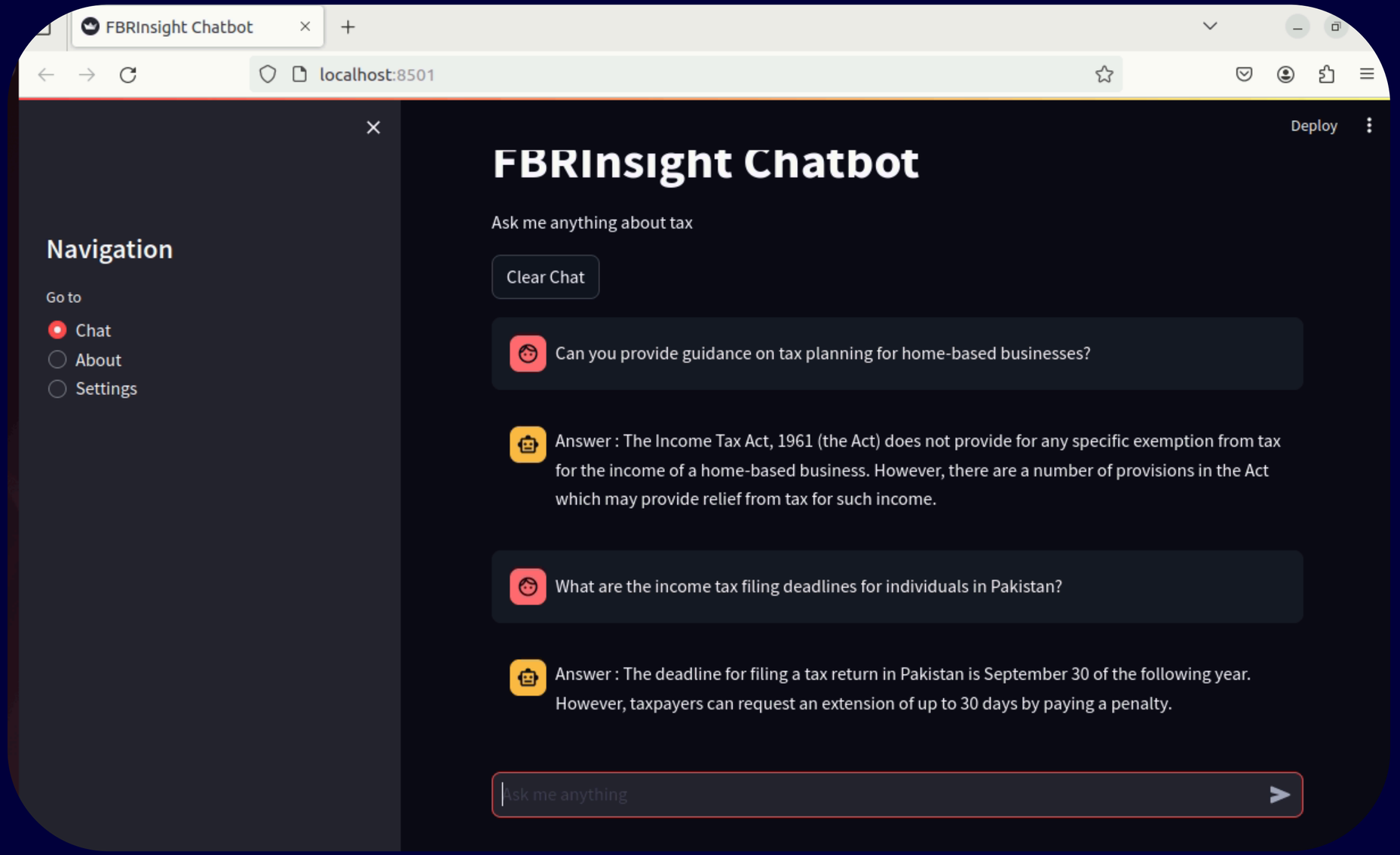
Types of Finetuning



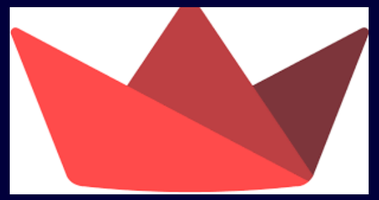
Training and Evaluation

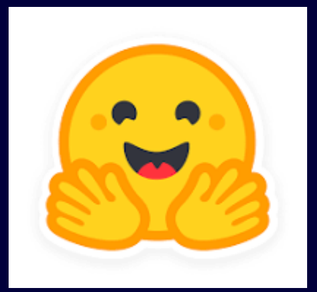


Final Product



Technology Stack





Future Work

1. Increase Training Data
2. Enhance Effeciency and Speed
3. On Premises Deployement
4. Mobile Application
5. Cross Platform Compatibilty


Questions?


Thank
you