AI-Powered PDF Assistant for Business

Leveraging LLMs for Extractive Question Answering and Dynamic Content Generation

LLM Use Case Overview

- **What are LLMs?**
  - Advanced AI systems for understanding and generating human language.

- **Potential in Document Processing:**
  - Automates information extraction and query answering.
  - Reduces manual effort and errors.

Convince the Customer or Business People

Value Proposition:

- Unique selling points of the AI assistant.

Real-World Examples:

- Success stories and case studies.

Address Pain Points:

- Show how it solves common challenges.

Demonstrate ROI:

- Highlight cost savings and productivity gains.

Focus on Business Value

- **Challenges Addressed:**
  - Manual processing inefficiencies.
  - Difficulty in information retrieval.

- **Key Benefits:**
  - **Productivity:** Automates tasks and speeds up retrieval.
  - **Accuracy:** Reduces errors and provides consistent responses.
  - **Cost Savings:** Lowers labor costs and training needs.
  - **Scalability:** Handles large volumes efficiently.

Development Process Overview

1. Data Preparation:

- **Collection:**
  - Gathered diverse and comprehensive datasets relevant to the document processing tasks.

- **Preprocessing:**
  - Cleaned and formatted the data to ensure consistency and quality.
  - Removed any irrelevant or noisy data to improve the model's performance.
  - Tokenized the text to convert it into a format suitable for model training.

2. Model Training:

- **Model Selection:**
  - Chose the `distilbert-base-uncased-distilled-squad` model for its efficiency and accuracy in question answering tasks.

- **Transfer Learning:**
  - Used transfer learning to leverage the pre-trained knowledge of the model and improve its performance on our tasks.

- **Training Process:**
  - Employed supervised learning techniques to train the model.
  - Split the dataset into training and validation sets to monitor the model's performance and prevent overfitting.

3. Evaluation:

- **Metrics:**
  - Evaluated the model using standard metrics like accuracy, precision, recall, and F1 score.

- **Iterative Improvements:**
  - Based on the evaluation results, made iterative improvements to the model by adjusting hyperparameters and retraining.

Extractive Question Answering

Training the Model:

- **Question-Answer Pairs:**
  - Created a dataset of question-answer pairs from the documents.
  - Ensured the questions covered a wide range of topics to enhance the model's generalizability.

- **Tokenization:**
  - Used tokenization to convert the text into tokens that the model can process.
  - Employed the DistilBERT tokenizer for consistent tokenization.

- **Training Steps:**
  - Fine-tuned the DistilBERT model on the question-answer pairs.
  - Employed loss functions to minimize the difference between predicted and actual answers.

Dynamic Content Generation

1. Dynamic Updates Based on KPIs:

- **Identifying KPIs:**
  - Extracted key performance indicators (KPIs) from the text using regular expressions.
  - Example KPIs: Length of the Nile River, Discharge rate.

- **Updating Content:**
  - Replaced old KPI values with new ones dynamically.
  - Ensured the context around the KPIs remained coherent and meaningful.

2. Implementation Steps:

- **Regex Patterns:**
  - Created regex patterns to identify and extract KPIs from the text.
  - Example: `'length_of_nile': r'6,650 kilometers'`

- **Updating Text:**
  - Used a function to replace old KPI values with updated ones in the text.
  - Highlighted changes to make updates transparent.

Summary and Q&A