1. Overall Approach

The chatbot implementation aims to create an interactive system capable of answering user queries using both predefined question-answer pairs and information extracted from a PDF document. The approach involves:

1. **Data Loading**:
   - Load question-answer pairs from a JSON file.
   - Extract and process text from a PDF document containing additional information.
2. **Text Processing**:
   - Use Natural Language Processing (NLP) techniques to tokenize and vectorize text for similarity matching.
3. **Similarity Matching**:
   - Employ cosine similarity to match user queries with predefined questions and extract relevant information from the PDF.
4. **Contextual Awareness**:
   - Maintain conversation history to provide context-aware responses, improving the relevance and coherence of interactions.
5. **Response Generation**:
   - Generate responses based on the similarity score from QA pairs and PDF content.
   - If no relevant information is found, prompt users to contact the business directly.

2. Frameworks/Libraries/Tools Used

a. Chainlit

- **Purpose**: Provides a framework for building and managing the chatbot interface.
- **Usage**: Handles user interactions, message sending, and receiving through its API.

b. PyMuPDF (fitz)

- **Purpose**: Extracts text content from PDF documents.
- **Usage**: Reads and processes PDF files to obtain the textual data used for answering questions.

c. NLTK (Natural Language Toolkit)

- **Purpose**: Provides tools for tokenizing text.
- **Usage**: Tokenizes text into sentences to facilitate text processing and similarity matching.

d. scikit-learn

- **Purpose**: Used for text vectorization and computing cosine similarity.
- **Usage**: `TfidfVectorizer` for vectorizing text and `cosine_similarity` for measuring text similarity.

e. JSON

- **Purpose**: Data format for storing and loading question-answer pairs.
- **Usage**: Provides a structured format for predefined questions and answers.

3. Problems Faced and Solutions

a. Problem: Inaccurate Similarity Matching

- **Issue**: The chatbot occasionally returned irrelevant information from the PDF.
- **Solution**: Adjusted the similarity threshold and improved text extraction methods to ensure only highly relevant matches are considered.

b. Problem: File Path Errors

- **Issue**: Encountered `FileNotFoundError` due to incorrect file paths.
- **Solution**: Verified and corrected file paths in the script to ensure accurate access to the JSON and PDF files.

c. Problem: Inconsistent PDF Content Extraction

- **Issue**: Extracted text from PDF was poorly formatted, affecting text processing.
- **Solution**: Enhanced text extraction methods and preprocessed PDF content for better readability.

d. Problem: NLTK Dependency

- **Issue**: Missing NLTK data required for tokenization.
- **Solution**: Added a script to download the necessary NLTK data during setup.

4. Future Scope

a. Enhanced Natural Language Understanding

- **Feature**: Integrate advanced NLP models like BERT or GPT to improve understanding and processing of user queries.

b. Multi-Language Support

- **Feature**: Expand the chatbot to support multiple languages, accommodating users from diverse linguistic backgrounds.

c. Contextual Awareness Enhancements

- **Feature**: Implement more sophisticated context management techniques to handle complex dialogue flows and improve the relevance of responses.

d. Integration with External APIs

- **Feature**: Connect the chatbot to external APIs to provide real-time information, such as stock prices, weather updates, or company news.

e. Improved PDF Processing

- **Feature**: Enhance PDF processing capabilities to handle complex document structures, including tables and images, for more accurate information retrieval.

f. User Feedback Mechanism

- **Feature**: Introduce a feedback system to collect user responses and continuously refine the chatbot based on user inputs.

Conclusion

This implementation outlines a foundational approach to building an interactive chatbot that utilises both predefined data and document-based information. Future enhancements can expand its capabilities, providing users with a more robust and versatile conversational experience.