Handling Hallucinations and Other Common Issues with LLMs

To handle hallucinations and improve the reliability of the LLM's responses, we can implement the following strategies:

a. **Use Retrieval-Augmented Generation (RAG):**
Instead of sending the entire relevant text to the LLM, we can use a vector database to store embeddings of the PDF content and retrieve the most relevant chunks. This helps focus the LLM on the most pertinent information.

b. **Implement Fact-Checking:**
After generating a response, we can use the LLM to generate a set of factual statements from the response. Then, we can verify these statements against the original PDF content using similarity search or regex patterns.

c. **Add Confidence Scores:**
We can ask the LLM to provide a confidence score for each part of its response. This can help users identify which parts of the answer might be less reliable.

d. **Use Prompt Engineering:**
We can optimize prompts that encourage the LLM to admit uncertainty and to stick closely to the provided information. For example:

- **Instructions:**
  1. Answer the question based only on the information provided in the resume excerpt.
  2. If you're unsure or if the information is not present, say so clearly.
  3. Do not make up or infer information that is not explicitly stated in the text.
  4. Provide a confidence score (0-100%) for your answer.

Your response should be in this format:

- **Answer:** [Your answer here]
- **Confidence:** [0-100%]
- **Reasoning:** [Explain your reasoning and cite specific parts of the text that support your answer]

e. **Implement a Feedback Loop:**
We can allow users to flag incorrect or hallucinated responses. And we use this feedback to fine-tune the LLM or adjust the retrieval process.

This is what i feel will help make the PDF chat application more robust and reliable, reducing hallucinations.