# Improving Large Language Model Responses for ISO9000 Queries

# User Question and Initial LLM Response

## Question:

What is ISO9000?

## LLM Initial Response:

I don't know.  
  
However, I can see that ISO 9001:2008 is mentioned in the provided text as a standard for quality management systems. The snippet mentions 'EN ISO 9001:2008' and states that it follows international standards such as this one.  
  
Snippet: 'The SHE Management Model provides a goal-setting framework, within which the Project Management Organisation (PMO) can develop detailed policies, standards, and procedures and incorporate them within their system. This can be applied from the start to the finish of the project. In addition to a framework, this guide provides exemplars of key procedures and documentation. The main sections of the guide are: status review; developing: goal setting, planning, organising, implementing, and measuring.'  
  
This suggests that ISO 9001 is related to quality management systems, but I don't know what specific aspects or requirements it entails.

## 1. Improving Context Retrieval and Response Generation

• Increase the similarity score threshold to around 0.3-0.4 to ensure more relevant content retrieval.  
• Limit the number of document chunks in context to prevent overwhelming the model and focus on the most relevant data.

## 2. Prompt Engineering for More Precise Answers

• Modify the prompt to request a concise, clear answer with specific references to the context.  
• Experiment with providing examples in the prompt to guide the model on specificity and format.

## 3. Improving Contextual Awareness of ISO Standards

• Supplement the vector store with foundational documents on ISO standards to provide context.  
• Pre-define reference sections for key terms like 'ISO 9000' to improve answer relevance.

## 4. Adjusting LLM Temperature and Response Mode

• Set temperature low (0.0–0.2) for consistent responses.  
• Experiment with output parsers that enforce concise answers.

## 5. Debugging and User Experience Enhancements

• Add checks to prevent model invocation when vector store is empty.  
• Use dynamic context logging to review context before sending it to the LLM, helping fine-tune retrieval settings.

## 6. Evaluation and Feedback Loop

• Regularly review responses and adjust settings based on accuracy.  
• Consider testing alternative models fine-tuned on QA tasks for potentially better responses.