# Quality Check and Code Improvement Suggestions for LLM Responses

This document summarises key suggestions to evaluate and improve the quality of responses generated by a Large Language Model (LLM) for specific project-related questions. It includes a quality check framework and recommended code adjustments for enhancing response relevance, specificity, clarity, and accuracy.

## 1. Quality Check Framework for LLM Responses

Evaluate each response based on the following criteria:

### Relevance

• Direct Answer: Ensure the response directly addresses the question without adding irrelevant information.  
• Comprehensiveness: Verify that all key points related to the question are covered and relevant.

### Specificity

• Level of Detail: Check that the response includes specific, actionable details rather than general statements.  
• Contextual Quotes: Ensure that key points are supported with direct quotes or paraphrasing from the provided context.

### Clarity

• Conciseness: The response should be brief but comprehensive, avoiding filler content.  
• Structured Response: For readability, use a bulleted or numbered list, especially when multiple points are discussed.

### Factual Accuracy

• Consistency with Context: Ensure information does not go beyond the context provided.  
• Accuracy of Content: Confirm that specific terms or risks are represented accurately based on context.

## 2. Code Adjustments to Improve Response Quality

### Update Prompt for Clarity and Structure

Modify the prompt template to guide the LLM towards providing structured, relevant responses:  
  
Template:  
  
You are an AI language model assistant. Your task is to provide a precise, structured answer to the question based ONLY on the provided context.  
  
Context: {context}  
  
Question: {question}  
  
Instructions:  
- Provide a direct answer using a bulleted or numbered list for readability.  
- Reference specific details from the context to support your response.  
- If the context does not provide enough information, state 'I don’t know,' but list any available relevant factors.

### Adjust Retrieval Settings

• Increase the similarity score threshold to ensure more relevant content retrieval.  
• Limit the number of chunks retrieved to focus on the most relevant information.

### Lower Temperature and Experiment with Different Parsers

Set the model temperature low (e.g., 0.1) to prioritise consistency and factual responses. Experiment with output parsers to enforce concise answers.

### Add Summary or Paraphrasing Step

Consider running a secondary step to summarise or paraphrase the response, enhancing clarity and conciseness.

## 3. Implement and Rerun Quality Check

1. Rerun the updated prompt and retrieval settings with the same question to assess improvements.  
2. Evaluate the new response based on relevance, specificity, clarity, and accuracy.  
3. Iterate as needed based on the assessment for further refinements.