Method	**Correct**
LLaMA alone	**10**
LLaMA + RAG	**3**

RAG vs No-RAG Observations:

1. **Accuracy and Contextual Relevance**

   - **RAG answers**: All 10 questions are answered accurately and concisely using the context from the documents. The answers directly reference the order details, revenue figures, conference info, etc.

   - **No-RAG answers**: Most answers fail to provide the requested information. The model frequently replies with disclaimers about not having access to real-time data or context, even though the information exists in your documents.

2. **Hallucinations**

   - **RAG**: All answers strictly based on context.
   - **No-RAG**: Much higher hallucination risk. For example, in Q1 it gives a completely wrong name ("Mian Ghulam Muhammad"), and for revenue questions, it provides explanations or historical guesses instead of the exact numbers.

3. **Benefit of RAG**

   - When **contextual documents are provided**, RAG ensures factual correctness by retrieving relevant information.

   - Without RAG, the model relies solely on its pre-trained knowledge and may give outdated, incomplete, or incorrect answers.

Conclusion:

RAG clearly outperforms standard LLM querying for document-based Q&A. The more detailed and structured your documents are (like emails, revenue reports, and event announcements), the better RAG performs, giving accurate and concise answers while drastically reducing hallucinations.