# Failure points of RAG system

by Dipanjan S., 5/1/24

Let's talk about key failure points in a Retrieval Augmented Generation (RAG) system as it is one of the most common questions I get. Check out this illustration which I adapted from a popular research paper which summarizes this in a nice visual! Failure points to be aware of when building RAG systems:  
  
1. Missing Content: Means the retrieval strategy didn't work and hence the LLM will reply I don't know or hallucinate  
  
2. Missed Top Ranked: The answer to the question is in your vector DB but did not rank highly enough to be returned to the user  
  
3. Not in Context - Documents with the answer were retrieved from the database but did not make it into the LLM context for generating an answer  
  
4. Not Extracted: Answer is present in the context, but the LLM failed to extract it. Typically, happens when there is noise or contradicting information in the context  
  
5. Wrong Format: Instruction involved generating information in a certain format such as a table or list and the LLM ignored it  
  
6. Incorrect Specificity: Happens when users are too vague about their question or answer is not specific enough or too specific and misses enough background information  
  
7.  Incomplete: Generated answer is correct but misses some of the information even though that information was in the context  
  
Some of these issues can be tackled by improving your retrieval strategy as I mentioned in my previous post, using a semantic cache, stronger LLMs and more grounded prompts and plugging in output parsers. However these are still active problems and if you have solved them feel free to share your thoughts!  
  
Referenced paper for the illustration and topics is, Seven Failure Points When Engineering a Retrieval Augmented Generation System, do check it out to dive into more details!

A diagram of a problem

Description automatically generated