Why Your RAG Doesn’t Work

breakdown of why many RAG solutions fail and how you can address these issues.

1

Retrieval Accuracy and Error Propagation

RAG start by retrieving relevant documents from a knowledge base. However, if the wrong documents are fetched, the errors carry forward, causing incorrect or irrelevant answers.

2

Handling Long Context Windows

Solutions often need to process long documents, such as reports or research papers. LLMs struggle when they have too much context, focusing only on the beginning or end and missing critical information in the middle.

3

Latency Issue

Users expect fast responses, but RAG, which must retrieve and process documents in real time, often suffer from slow response times due to high computational demands.

4

Inconsistent Information

Based on your strategy RAG solution may pull data from several documents that have conflicting information. Without a way to resolve these discrepancies, the generated answers can be confusing.

5

Cost and Scalability

Running RAG at scale can be expensive and resource intensive due to the heavy computation involved in retrieval, ranking, and generation.

Some Possible Solutions:

1. Implement high accuracy retrieval strategies using multi query, multi or cross vector searches on index and agents.
2. Manage context windows by chunking long documents smartly and additional metadata matching.
3. Focus on reducing latency by using semantic caching and parallel processing techniques to speed up retrieval and generation.
4. Ensure consistency across multiple documents and flag contradictions.
5. Implement batch processing for similar queries to reduce cost and scalability to maintain efficient operations at scale