Retrieval-Augmented Generation

Computing

The pipeline that lets a model answer from your documents.

Two phases: index once, then retrieve and generate on every question.

1. **Chunk**
   Split documents into passages — commonly 200–800 tokens, with a little overlap.

2. **Embed and store**
   Vector for each chunk, kept in a vector database with its source.

3. **Retrieve**
   Embed the question, fetch the nearest chunks.

4. **Rerank**
   Optionally re-score the top 50 down to the top 5 with a stronger model.

5. **Generate**
   Put the chunks in the prompt and ask the model to answer only from them.

WHERE IT GOES WRONG

Chunks too large — the relevant sentence is diluted
Chunks too small — the context is missing
No citations — the answer cannot be checked

> **Ask the model to say when the retrieved text does not contain the answer. Without that instruction it will guess.**