**MODEL RETRIEVAL**

It is helps models to updated with latest information, without the new latest information retrieval model will always give old results.

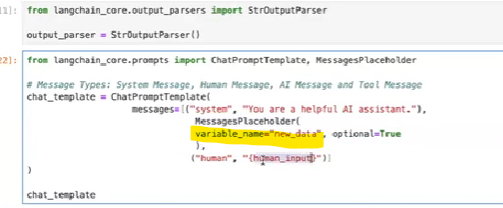
The model will give hallucinated answer.

For overcome this issue RAG is used here.

On RAG we attached a knowledge base the model from that knowledge base all the latest info will be retrieved.

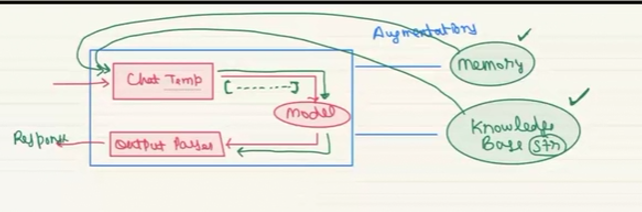
We will give place holder to hold the latest info and that will be retrieved later

**SOLUTION 1:**



In the above approach we have one problem that is when a size of knowledge base increases then the prompt goes to model will be huge

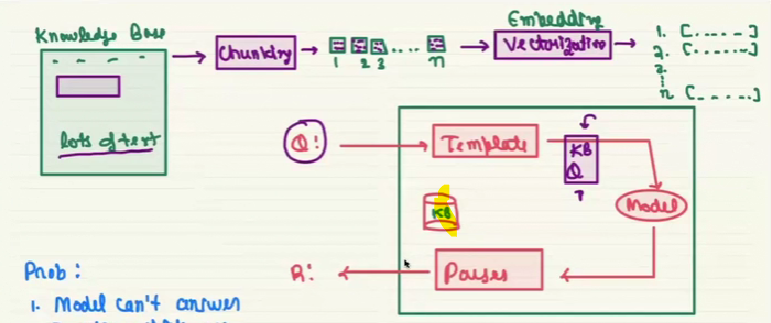
EX: Assume knowledge base containing the 100 various document each have 1000+ pages



Model will not process whole knowledge base information because it has limited constant window that will take some size of data from knowledge base instead of taking whole data.

To over come this issue **chunking** is used.

On chunking we will pass needed part of file to give correct answer to the user question by we converting whole text knowledge base values to numeric vectors and that vectors are will be stored in a vector DB.



User input also are gets converted into numeric representation and that numeric user input are passed to vector DB to find the correct answer for the user query from those collection of chunks. To identify the similar file from vector DB along with converted user input we have use another method called **cosine similarity** (works well for high dim data)**.** And that founded chunk are the only info which are passed to model.

Below is the image which elaborates the above theory process.

