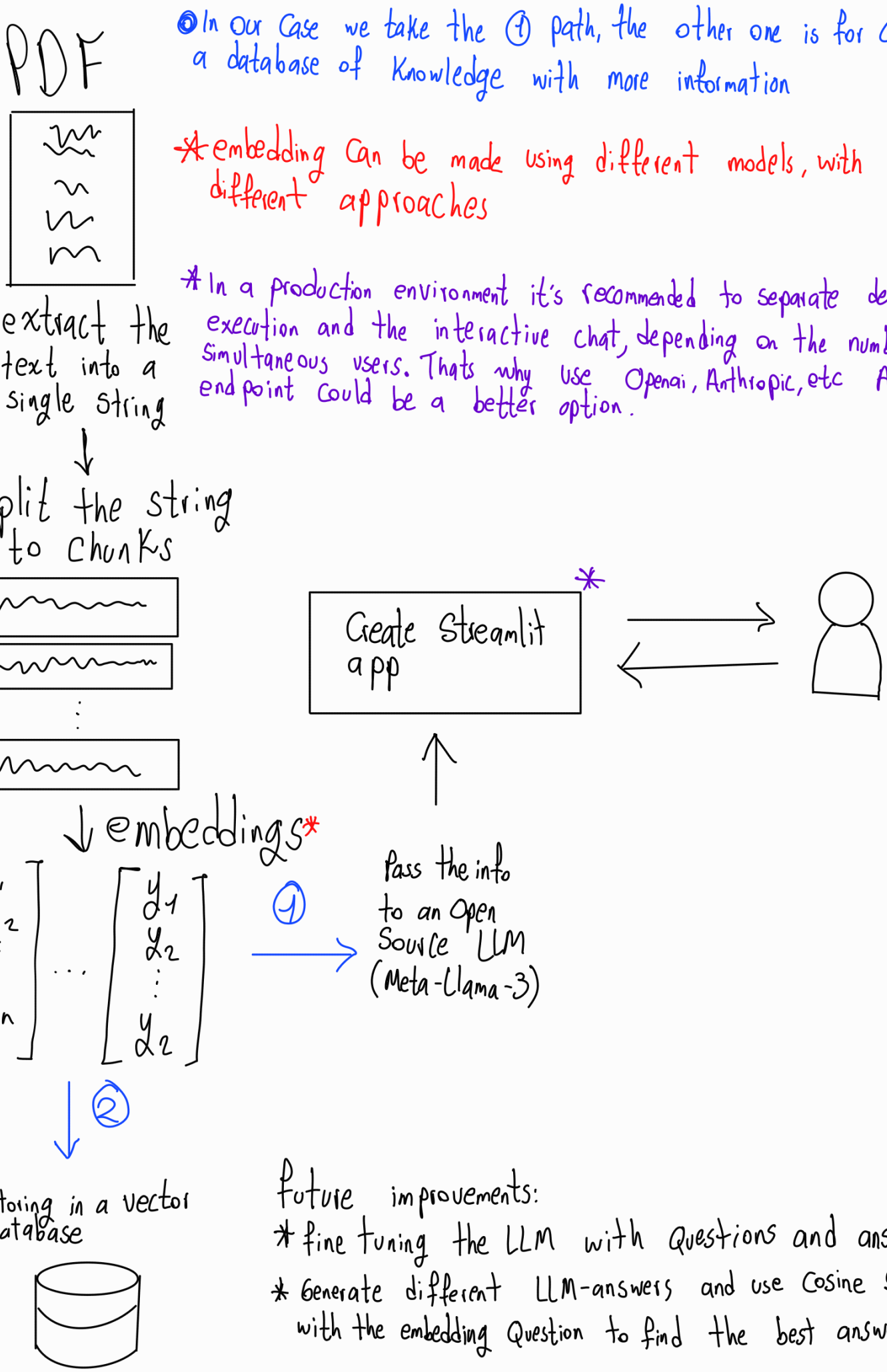
PDF
In our case we take the ① path, the other one is for create a database of knowledge with more information
*embedding can be made using different models, with different approaches
* In a production environment it's recommended to separate de LLM execution and the interactive chat, depending on the number of simultaneous users. Thats why use Openai, Anthropic, etc API endpoint could be a better option.
extract the text into a single string
split the string into chunks
Create Streamlit app
*
embeddings*
Pass the info to an Open Source LLM (Meta-Llama-3)
①
②
Storing in a vector database
future improvements:
* fine tuning the LLM with Questions and answers
* Generate different LLM-answers and use Cosine Similarity with the embedding Question to find the best answer.