Answers to questions

1 What type of questions worked ? What questions didn't work?
Questions on the text itself worked
Questions like to help to solve the exercise or even to count the number of words didn't work
I used a similarity_search technique that is based on the semantic similarity in the question

2 What improvements should be made to the service to support the non-answered questions?
I'm not an AI expert but I imagine to build a system that combines multiple techniques more matching a large spectrum of questions and that will analyze the question before to know to which technique to use

3 Design and architecture to support parallel loading of 100M documents

First I have to say, in my code i didn't optimize it.
All the first instructions like loading the doc, split it, creates the vectors and store them in Chroma DB needs to be separated in a kind of init process when i start the docker and it will be in cache. This will be done only once.
Then my endpoint will only answer the question.
For now on each question all the instructions are repeated.

Now, concerning the question about the 100M documents, I guess I need to build an infrastructure that would be able to process all the documents in a distributed way.
There are lots of technologies to distribute processes like Spark, for example.
So maybe the idea is to put all the docs on something like S3, then Spark will get them and distribute all the processes. The embeddings will be stored on a vectorial DB like Milvus that it will be indexed with an algo like HNSW. (i didn't know about this algo and thanks about the exercise that gave me the opportunity to learn)
The AI model will be able to use this DB to answer