Dany's Questions

1 - What is fine-tuning RAG?

Ans:

Basically Fine-tuning in NLP is like teaching an already educated, Chatbot/RAG assistant some new, specific skills. OR to train it on your data like we are doing in your project.

Let me clarify what we are doing in our project:

We are using GPT-3.5 turbo for now (will replace it with GPT-4 turbo or the latest version), now this LLM is not trained on our data, for example, it does not know **"who is Dany, and Sarfaraz"**, but to make it possible we are using the concept of **RAG (with vector database)** through which we are also training it on our own data.

2 - Secondly, I wanted to know if the data had to be dated so that the model could know which data to prioritize.

Ans:

Yes, that's not a big deal. if the data is dated, or updated, we will just have to give some commands in the system message to the LLM to give the response accordingly.

Such as "Hey chat, if you have updated data, please do prioritize it, else answer with old data"

3 - I also wanted to know how the model works. Does it really integrate and reason with the data I provide, or when I submit a query does it just copy/paste the sources it has in its possession?

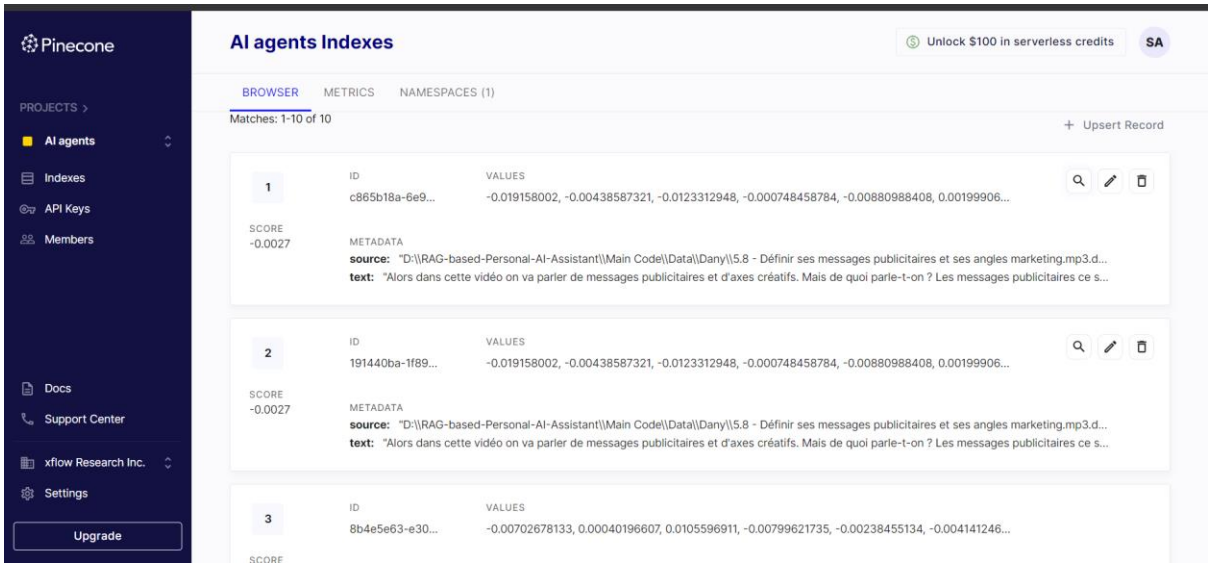
Ans:

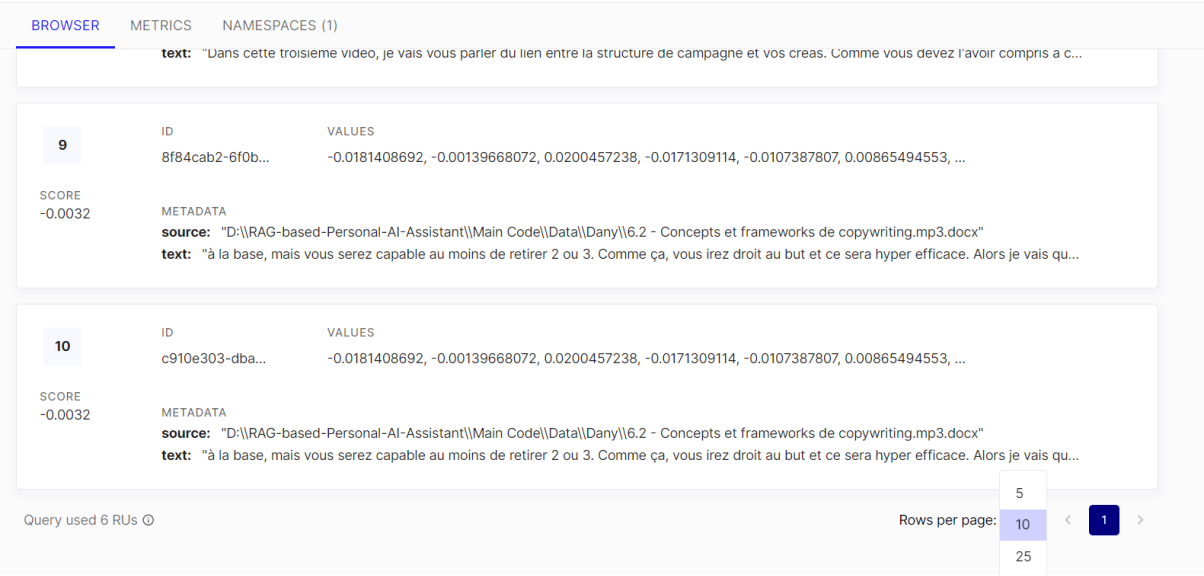
It does not just copy and paste, your data is converted into embeddings and then stored in the pinecone vector database.

when we query it, then the integrated LLM provides us with the response by refining the data in a structured format.

See in screen shoot:

Pinecone vector Database (where your data is loaded into embeddings form when then retrieved by LLM based on similarity search).





4 - Lastly, I also have blog article content with images. Do you think I should stick to text-only content, or is it possible to provide him with hybrid documents? And is it also possible to provide him with data in English, or do they have to be in French only?

Ans:

Yes, we can train RAG on hybrid documents or blog articles with images as well. (but I will have to do some research). And as of my experience, it's possible.

~Thank you!