What I Did and Why I Did It

As you may know, the GPT-3.5 API, offered by OpenAI, now comes with associated costs. It is no longer a free API, which has led me to explore alternative solutions for building models that do not incur such charges.

Initially, I considered using several other free API cloud-hosted models, such as Hugging Face's free offerings, BigScience BLOOM, and Cohere API. However, I encountered significant challenges with these options. Despite spending several hours troubleshooting and fixing bugs, To complete the task in the limited time period I found that these models are not as straightforward or easy to use as GPT-3.5. The user experience was not as smooth, and the documentation and implementation were not as accessible.

Additionally, Hugging Face offers free LLM models that work well for many use cases. However, when I tried to implement Retrieval-Augmented Generation (RAG) with these models, they started to download the models locally. This posed another challenge because certain models, such as Mister and others, require more than 10 GB of storage space. While this is manageable locally, it becomes a problem when deploying on free instances, as these typically have limited storage and computing capabilities.

Considering these limitations, I concluded that it is better to explore GPT-3.5 alternatives (cloud-hosted models with free APIs) for a smoother and more efficient development experience. Some options may still offer the capabilities needed for various tasks without the storage and deployment issues faced with other models.