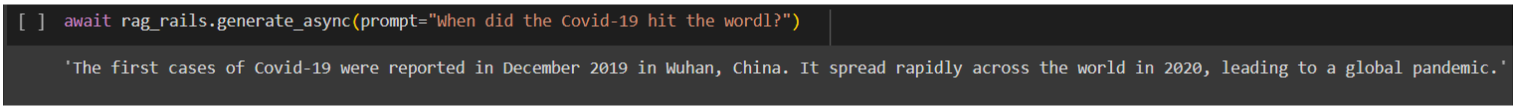
## *RAG Inference*:

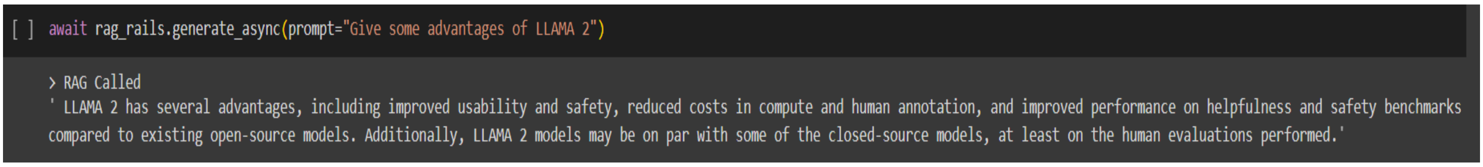
After building the whole pipeline, we are querying the model. So here as you can see, we have asked the model a very generic question which is not related the doc. Hence the RAG pipeline is not called here, and the model has answered the question.



```
[ ]  await rag_rails.generate_async(prompt="When did the Covid-19 hit the wordl?")

     'The first cases of Covid-19 were reported in December 2019 in Wuhan, China. It spread rapidly across the world in 2020, leading to a global pandemic.'
```

Now we have asked the question which is related to the document, which is provided, Now you can see the pipeline is automatically called and the answer generated through it is perfect.



```
[ ]  await rag_rails.generate_async(prompt="Give some advantages of LLAMA 2")

     > RAG Called
     ' LLAMA 2 has several advantages, including improved usability and safety, reduced costs in compute and human annotation, and improved performance on helpfulness and safety benchmarks
     compared to existing open-source models. Additionally, LLAMA 2 models may be on par with some of the closed-source models, at least on the human evaluations performed.'
```