04/Nov/2023

Report

Main Accomplishments:

1. Run a pipline to produce a story 01/NOV/2023

* Successfully made a pipline using Autogen framework to generate story.
* Although, generally I need to pay about £1 to do a simple experiment and the amount of estimated cost of this project is much higher than what I can afford. (Also, it seems that the more agents and prompt I uses the payment would grow exponentially which is much higher than this estimation)

Next goal:

* Try some open source free models for this project
* Run the pipeline that generates deceptive story(in general)

1. Find a method to run LLaVA manually 04/NOV/2023

* The LLaVA use large and heavy for my laptop to run and when it finally run it was too unstable
* I have managed to run the model on google Coolab which is very stable
* Although, to use teachable agent feature of the autogen framework, I think there are some limitations due to the limited use of the server.

Next goal:

* Modify the pipline to be executable using LLaVA model
* Modify the agent prompts and setting to implement the TattleTale model(<https://icaps22.icaps-conference.org/workshops/SPARK/papers/spark2022_paper_2.pdf>)
* Add deception agent to modify the story to be deceptive

1. Fixed Error with bitsandbytes:

* It turns out that bitsandbytes library do not work on window:

Link: <https://github.com/TimDettmers/bitsandbytes/issues/748>

1. Error: KeyError: “LLaVA”

* Need to pretrain the model on the local machine

1. Problem!: RAM required over 12G

* My laptop kept failing to pretrain the model because of the amount of RAM needed -need more than 8Gb
* Colab also failed -need more than 12Gb

RESOLVED!: Colab pro with 51Gb barely managed to pretain the model of 7b model (llava-v1.5-7b)

* It seems model with higher number of parameters is much difficult to run

1. Problem!: Not enough terminal for google colab

* Use colab-xterm: similar to linux

1. Comparison with 7b and 13b parameter:

* The accuracy of the answering skill differs greater than my expectation. 7b tends to hallucinates and makes out of context response much greater than any other LLM(gpt-3,gpt-4, LLaVA 1.5 13b)
* Although this can be overcome by using gpt-4 as an agent manger(judge) checking LLaVa’s response

1. Using a Server on LLaVA is quite unstable:

* When the google server goes through high traffic the model produces this message:

\*\*NETWORK ERROR DUE TO HIGH TRAFFIC. PLEASE REGENERATE OR REFRESH THIS PAGE.\*\*

* The next agent receive this message, then produce inaccurate response: leading the pipeline to be unable to run
* Here is an experiment of giving a goal to plot a graph using a data online: (I found this experiment useful because it uses all of the features of LLaVA and the Autogen framework to accomplish a certain goal)

[experiment on features of LLaVA and stablilty of the framework.txt](experiment%20on%20features%20of%20LLaVA%20and%20stablilty%20of%20the%20framework.txt)

**Conclusion:**

From this week experiments with trails and errors. I have found out that the LLaVA machine requires large RAM memory about 50Gb and

In order for this condition to be made google colab was my method.

However, when the google colab servers are going through high traffic the pipeline was very unstable.

So I would need to run the experiment on private server such as “Replicate” or other platform that provides gpu processing power online.

Although not much have been achieved on this week. I have found out that LLaVA has some potential in other field that gpt-4 lacks of.

* It is multimodel so it can process image and videos as well. Which in my experience, much accurate and reliable than gpt-4
* Using many models which are specified in certain fields is very effective on Autogen framework