Software Engineering – Homework 1

Uritu Andra-Ioana, 1241EB

Model	Temp	Response	Speed
distilbert/distilgpt2	0.2	Today I learned how to run a local LLM because it was so easy. I’ve been using the tool for over 20 years and have always wanted to learn more about what is possible with this new feature, but now that we are able to use our own tools in order to make sure	~7–10s
	0.7	Today I learned how to run a local LLM because it was really easy, and what is interesting about the project. In this article we will look at one of the most common methods for running our application in your own language: compile (or use) many different languages into an executable or file called	
	1.2	Today I learned how to run a local LLM because the process is very slow. It turns out it has worked really well when you start building an architecture and then there's just one problem (and maybe we've not solved yet), so if everything was fine, take some time until next month for that	
openai-community/gpt2	0.2	Today I learned how to run a local LLM because it was the only way. I have been doing this for over 20 years now and my wife is still learning about what makes an effective, efficient system work well in practice. It's not just that you can't do anything with your code anymore	~10–15s
	0.7	Today I learned how to run a local LLM because it was easy and had great features. So, when you're in the middle of your project with no one around that time saying "I'm gonna write some code", then this is what we do: 1 2 3 4 5 6 7 8	
	1.2	Today I learned how to run a local LLM because my mother was afraid of the disease, but then it went away and that scared all."	

		"You can do other things. The main thing about running in this country is everything else makes you feel bad when nothing gets done—you never know if	
Qwen/Qwen2-1.5B-Instruct	0.2	Today I learned how to run a local LLM because of the help of my friend, who is an expert in this field. He explained everything step by step and it was really helpful for me. We also discussed some other interesting topics related to machine learning and artificial intelligence that we found fascinating. It was	~20–35s
	0.7	Today I learned how to run a local LLM because of the following: * I'm using my MacBook Pro * I have a Google Colab notebook running in it The instructions are as follows: Go to colab -> Run Code Cell (which is in the left) and click on "Submit	
	1.2	Today I learned how to run a local LLM because we were working remotely during the lockdown and in order to communicate with each other in our projects, we needed to create a communication tool. The best idea I found was to make a live link between two computers in your network, something similar to ssh.	

Conclusions:

Temperature effects:

- 0.2 = most focused but sometimes dull or weirdly literal (GPT-2 had strange specifics even at low temp).
- 0.7 = best balance: distilgpt2 stayed reasonable and GPT-2 showed some weird parts; Qwen produced structured steps.
- 1.2 → most creative but drifted a lot (GPT-2 especially) and sometimes invented irrelevant details.

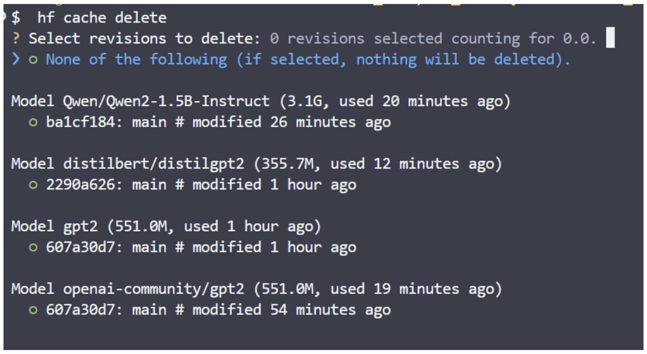
Model differences:

- distilbert/distilgpt2 (≈82M): small and fast; surprisingly good at 0.2–0.7; starts to ramble nonsense at 1.2.
- openai-community/gpt2 (≈124M): bigger but not consistently better, more odd deviations than distilbert

- Qwen2-1.5B-Instruct: slowest but most structured and “helpful,” especially at 0.2–0.7

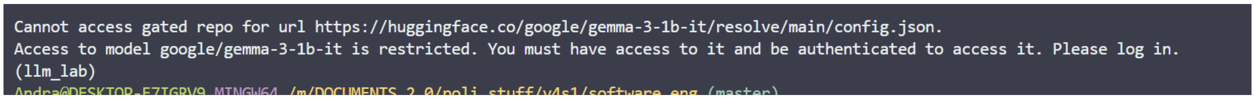
What I found surprising:

- Even small models take a lot of space. for example, microsoft/phi-2 used around 5.6 GB.
    - I did not make a screenshot of microsoft/phi cache size before erasing it but even Qwen which has 1.5B takes up over 3gb of space



```
$  hf cache delete
? Select revisions to delete: 0 revisions selected counting for 0.0.
> o None of the following (if selected, nothing will be deleted).

Model Qwen/Qwen2-1.5B-Instruct (3.1G, used 20 minutes ago)
  o ba1cf184: main # modified 26 minutes ago

Model distilbert/distilgpt2 (355.7M, used 12 minutes ago)
  o 2290a626: main # modified 1 hour ago

Model gpt2 (551.0M, used 1 hour ago)
  o 607a30d7: main # modified 1 hour ago

Model openai-community/gpt2 (551.0M, used 19 minutes ago)
  o 607a30d7: main # modified 54 minutes ago
```

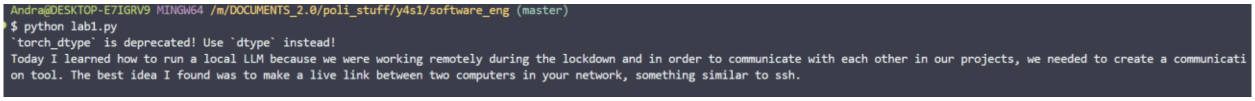
- The bigger GPT-2 model did not always perform better than the smaller distilbert one.
- Some LLMs cannot be used without logging in:



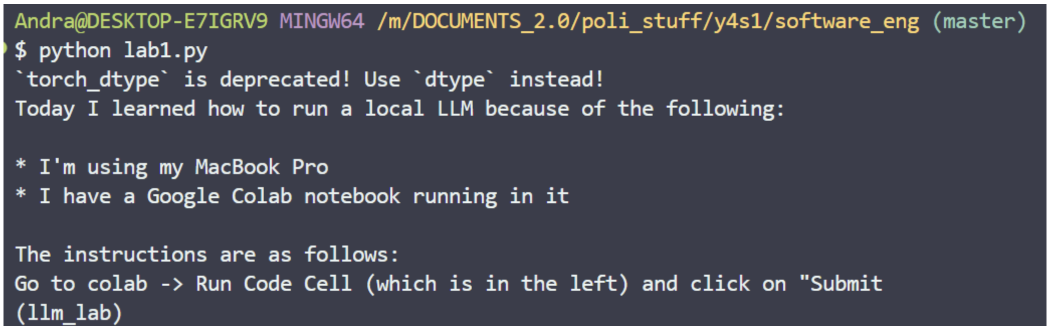
```
Cannot access gated repo for url https://huggingface.co/google/gemma-3-1b-it/resolve/main/config.json.
Access to model google/gemma-3-1b-it is restricted. You must have access to it and be authenticated to access it. Please log in.
(llm_lab)
Andra@DESKTOP-E7IGRV9 MINGW64 /m/DOCUMENTS_2.0/poli_stuff/y4s1/software_eng (master)
```

Note:

For this homework I chose to run the models locally just to see how it works. Here are some screenshots of the answers in bash:



```
Andra@DESKTOP-E7IGRV9 MINGW64 /m/DOCUMENTS_2.0/poli_stuff/y4s1/software_eng (master)
$ python lab1.py
`torch_dtype` is deprecated! Use `dtype` instead!
Today I learned how to run a local LLM because we were working remotely during the lockdown and in order to communicate with each other in our projects, we needed to create a communicati
on tool. The best idea I found was to make a live link between two computers in your network, something similar to ssh.
```



```
Andra@DESKTOP-E7IGRV9 MINGW64 /m/DOCUMENTS_2.0/poli_stuff/y4s1/software_eng (master)
$ python lab1.py
`torch_dtype` is deprecated! Use `dtype` instead!
Today I learned how to run a local LLM because of the following:

* I'm using my MacBook Pro
* I have a Google Colab notebook running in it

The instructions are as follows:
Go to colab -> Run Code Cell (which is in the left) and click on "Submit
(llm_lab)
```