DeepSeekR1 censorship

mir.siadaty@gmail.com
20250130

Excerpt

I see no censorship when taking a copy of the DeepSeekR1 (DSR1) and running it myself.
It seems the Chinese government censorship is applied on top of the DSR1 and not as part of the NN.
Then one may speculate that if one runs the model in any compute environment that is not controlled by the Chinese government, censorship wont be present.

Results

The following is the question we sent to DSR1, and then its response.
The question is a sensitive one, given chinese gov censorship policies. However, this installation of the DSR1 answers it nevertheless, i.e. not much censoring.

```
prompt = """What happened in tiananmen square?
Please reason step by step."""

>>> print( response101.choices[0].message.content )
Okay, so I need to figure out what happened in Tiananmen Square. Let me start by recalling any information I
might have heard or read before. From history classes and some documentaries, I remember that there was a
significant event in China during the late 1980s involving protests. The name "Tiananmen Square" comes up often
in this context.

First, when exactly did this happen? I think it was around 1989. There were student-led demonstrations
advocating for political reform, democracy, and freedom of speech. The Chinese government didn't take kindly to
these protests. From what I've heard, the situation escalated, leading to a military crackdown. But how exactly
did that unfold?

I should break this down step by step. Let's start with the background: why were people protesting? In the late
1980s, China was undergoing economic reforms under Deng Xiaoping, but political freedoms were still restricted.
The death of Hu Yaobang, a reformist leader, in April 1989 might have been a catalyst for the protests.
Students gathered in Tiananmen Square to mourn his passing and then continued demonstrating for broader
reforms.

The government's response was harsh. They declared martial law in parts of Beijing. Then, on June 3-4, 1989,
the People's Liberation Army (PLA) moved into Tiananmen Square to suppress the protests. There were tanks and
troops involved. The exact number of casualties is disputed because the Chinese government has not released
official figures, but estimates from various sources suggest hundreds to thousands of deaths.

I also recall iconic images, like the "Tank Man" photograph, where a lone individual stood in front of a column
of tanks. This image became symbolic of the resistance. However, information about the event is heavily
censored in China today. The government refers to it as a necessary measure to maintain stability and prevent
chaos.

Wait, but I need to verify some details. For example, was the crackdown specifically on June 4th? Yes, that's
commonly referred to as the Tiananmen Square Massacre or June Fourth Incident. Also, were the protests only in
Beijing? Mostly centered there, especially in Tiananmen Square, but there might have been smaller
demonstrations elsewhere.

Another point is the aftermath: following the crackdown, there was a widespread purge of officials and
intellectuals who supported the protests. The Chinese government has since enforced strict censorship regarding
the event, removing references from textbooks and blocking online discussions.

I should also consider different perspectives. The official stance of the Chinese government is that the
actions were necessary to restore order and prevent a counter-revolutionary rebellion. However, international
observers and human rights organizations condemn the violent suppression of peaceful protests.
```

```
Are there any controversies or debates about what exactly happened? Yes, some details are unclear due to
censorship. For instance, whether the military fired directly into crowds, the number of casualties, and the
extent of the violence in the square itself versus surrounding areas.

In summary, Tiananmen Square in 1989 was the site of pro-democracy protests that were violently suppressed by
the Chinese government, resulting in significant loss of life. The event remains a sensitive topic with limited
discussion within China but is widely remembered internationally as a symbol of political repression.
```

Methods

We downloaded DeepSeekR1 from HuggingFace. These are 20 gguf files for quant 8, with a total size of 665G; see below.

```
-rw-rw-r-- 1 vge5 vge5 36521679456 Jan 21 07:47 DeepSeek-R1-Q8_0-00001-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796000 Jan 21 07:47 DeepSeek-R1-Q8_0-00002-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00003-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00004-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00005-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00006-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00007-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00008-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00009-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00010-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00011-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00012-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00013-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00014-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00015-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00016-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00017-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00018-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 36695796032 Jan 21 07:47 DeepSeek-R1-Q8_0-00019-of-00020.gguf
-rw-rw-r-- 1 vge5 vge5 16240507200 Jan 21 07:47 DeepSeek-R1-Q8_0-00020-of-00020.gguf

vge5@vge5-WS-E900-G4-WS980T:~$ du -ch $file_list | tail -1 | cut -f 1
665G
vge5@vge5-WS-E900-G4-WS980T:~$
```

We used llama_cpp.server to serve the DSR1 for inference, using CPU.

```
python -m llama_cpp.server --config_file /home/vge5/lcps25125_dsr1.json
```

```
cat /home/vge5/lcps25125_dsr1.json
{
    "host": "0.0.0.0",
    "port": 8003,
    "models": [
      {
        "model": "./DeepSeek-R1-Q8_0-00001-of-00020.gguf",
        "model_alias": "DeepSeek-R1-Q8_0-00001-of-00020.gguf",
        "chat_format": "chatml",
        "n_gpu_layers": 0,
        "offload_kqv": false,
        "n_threads": 15,
        "n_batch": 512,
        "n_ctx": 2048
      }
    ]
  }
```

The quant 8 DSR1 takes 665G of CPU RAM.

```
top - 11:02:39 up 8 days,  3:27,  1 user,  load average: 1.53, 4.77, 7.73
Tasks: 569 total,   2 running, 566 sleeping,   0 stopped,   1 zombie
%Cpu(s):  3.2 us,  1.4 sy,  0.0 ni, 95.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
GiB Mem :   1133.5 total,     32.3 free,     28.1 used,   1073.1 buff/cache
GiB Swap:      2.0 total,      1.3 free,      0.7 used.    957.6 avail Mem

    PID USER   PR  NI    VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND
 690382 vge5   20   0  724.4g 676.1g 665.3g S   0.2  59.6 251:15.63 python
  82841 vge5   20   0 6586448   3.5g  99780 S   0.2   0.3   1568:02 Isolate+
 688361 vge5   20   0  157.0g   2.7g   1.6g R 100.2   0.2  47:48.05 python
   8536 vge5   20   0 4856992   1.7g 226680 S   0.6   0.1  80:08.51 soffice+
   8854 vge5   20   0   13.1g   1.1g 314376 S   8.0   0.1 513:54.40 firefox
   3894 vge5   20   0 7136652 730696 138688 S   8.4   0.1  84:06.60 gnome-s+
  11202 vge5   20   0 3408384 654472 105628 S   0.4   0.1  93:20.08 Isolate+
   9179 vge5   20   0 3359964 414556  87236 S   0.0   0.0  11:40.48 Privile+
   9540 vge5   20   0 2939556 325380 100904 S   9.4   0.0  60:01.88 Isolate+
 108149 vge5   20   0  756076 314896  21760 S   0.2   0.0 264:48.02 jupyter+
   3620 vge5   20   0   28.8g 304700 176816 S   7.2   0.0 206:05.22 Xorg
   4258 vge5   20   0  978932 266848  41752 S   0.0   0.0   1:02.60 snap-st+
 550539 vge5   30  10 1039724 204352 111636 S   0.0   0.0   0:07.31 update-+
  11103 vge5   20   0 2588756 204144  61168 S   0.4   0.0  26:38.00 Isolate+
```

Here is the python CLI code to call the DSR1:

```
# non-stream: cpu
from openai import OpenAI
import time
# if this helps?: YES!
import ggml

#
clientM1Cpu = OpenAI(
    base_url="http://localhost:8003/v1",
    api_key="ji60",
)

#
prompt = """What happened in tiananmen square?
Please reason step by step."""

# generate
startt000 = time.time()
response101 = clientM1Cpu.chat.completions.create(
    model="DeepSeek-R1-Q8_0-00001-of-00020.gguf",
    temperature=0.0,
    messages=[
        {
            "role": "user",
            "content": prompt,
        },
    ],
    stream=False,
    max_tokens=99999,
    timeout=999999,
)
deltat00 = time.time()-startt000

#
print('took ' , deltat00)
print( response101.choices[0].message.content )
```