# Report

## 1

- The backbone model (OPT-2.7b) size: 5.0G
- Delta checkpoint size: 31M

## 2

- The largest size of the OPT model I can use with delta tuning is OPT-2.7b, run time GPU memory:

```
ubuntu@ml-ubuntu20-04-desktop-v2-6-32gb-25m ~/B/e/L4_prompt_delta_tuning (main)> nvidia-smi
Sun Jul 10 09:57:45 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 510.73.05    Driver Version: 510.73.05    CUDA Version: 11.6     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  NVIDIA GeForce ...  Off  | 00000000:06:00.0 Off |                  N/A |
| 55%   60C    P2   247W / 350W |  21813MiB / 24576MiB |     46%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A      1149      G   /usr/lib/xorg/Xorg                  4MiB |
|    0   N/A  N/A     57679      C   python3                         21805MiB |
+-----------------------------------------------------------------------------+
```

- without delta tuning, it will raise error:

```
 RuntimeError: CUDA out of memory. Tried to allocate
100.00 MiB (GPU 0; 23.70 GiB total capacity; 21.17 GiB
already allocated; 24.81 MiB free; 22.51 GiB reserved
in total by PyTorch) If reserved memory is >>
allocated memory try setting max_split_size_mb to
avoid fragmentation.  See documentation for Memory
Management and PYTORCH_CUDA_ALLOC_CONF
```

## 3

- GPU status with delta tuning(OPT-350m):

```
ubuntu@ml-ubuntu20-04-desktop-v2-6-32gb-25m ~> nvidia-smi
Sat Jul  9 07:41:45 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 510.73.05    Driver Version: 510.73.05    CUDA Version: 11.6     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  NVIDIA GeForce ...  Off  | 00000000:06:00.0 Off |                  N/A |
| 41%   38C    P2   128W / 350W |   6693MiB / 24576MiB |     19%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A      1119      G   /usr/lib/xorg/Xorg                  4MiB |
|    0   N/A  N/A     33503      C   python3                          6685MiB |
+-----------------------------------------------------------------------------+
```

- GPU status without delta tuning(OPT-350m):

```
ubuntu@ml-ubuntu20-04-desktop-v2-6-32gb-25m ~> nvidia-smi
Sat Jul  9 07:43:15 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 510.73.05    Driver Version: 510.73.05    CUDA Version: 11.6     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  NVIDIA GeForce ...  Off  | 00000000:06:00.0 Off |                  N/A |
| 41%   44C    P2   145W / 350W |  10851MiB / 24576MiB |     24%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A      1119      G   /usr/lib/xorg/Xorg                  4MiB |
|    0   N/A  N/A     34413      C   python3                         10843MiB |
+-----------------------------------------------------------------------------+
```

# 4

- Some questions and the model's answers:

| QUESTION | OPT-350M | OPT-2.7B |
|---|---|---|
| How many stars are there on the American flag? | 100 | 50 |
| How far is the earth from the sun? | the moon | 93 million miles |
| What is the use of plant stems? | for food | to carry food and water to the roots |
| What is CPU? | CPU is a computer program that runs on the computer | central processing unit |
| Where is China? | China's financial hub | East Asia |
| What is Github? | a site where people write code | a website where you can host code |
| Who is the founder of GitHub? | a person who owns a company | Chris Wanstrath |
| Who is the author of Das Kapital? | Hans Christian Andersen | Karl Marx |

- Obviously, the performance of OPT-2.7b is much better than OPT-350m.