github: <https://github.sec.samsung.net/haoyan-yang>

ML platform: <https://mlp.sjc.sr-cloud.com>

ML Q&A:

How to upload the file in the VsCode to the GitHub every time

git init

git add [file name]

git commit -m "xxx"

git branch -M main

git remote add origin [xxx]

git push -u origin main

username: haoyan-yang

password: 2313478Yhy!

Saving path: gpfs-volume

Check the gpu status: watch -n1 nvidia-smi

cd /gpfs-volume; python3 new\_monitor.py

If want to show the terminal, press ctrl + ~ (left top)

Install the anaconda in the server:

wget <https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh>; bash Miniconda3-latest-Linux-x86\_64.sh; source ~/.bashrc

conda create -n myenv python=3.10; conda activate myenv; conda install pip

python -m pip install .

python -m pip install flash-attn --no-build-isolation

pip install -r requirements.txt

pip install -e .

huggingface-cli login --token "${your\_access\_token}"

hf\_GXXiGUUEiXpTIWhkjaMqmvUOEVjrttzlKt

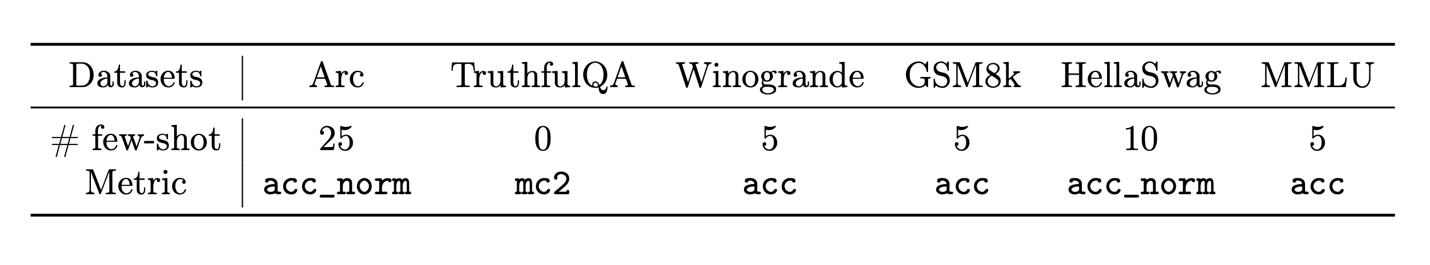
pip install torch==2.1.2

pip install transformers == 4.38.0

pip install accelerate==0.27.2

evaluation code

accelerate launch lm\_eval --model hf --model\_args pretrained=/gpfs-volume/SPIN/outputs/llama-tiny-20k/iter0 --tasks arc\_easy,arc\_challenge --device cuda --batch\_size 8 --num\_fewshot 25



look up the loss

tensorboard --logdir=xxx

**SPACE**

ssh -D 12345 -f -C -N vmwspacejmp7sc

accelerate launch --config\_file multi\_gpu.yaml --num\_processes=2 --main\_process\_port 29500 sft.py

conda activate spinenv; cd /user-volume/SPIN; bash scripts/generate.sh

pip install transformers==4.38.2

pip install accelerate==0.27.2

check the size of files: du -sh file\_path

for the evaluation of WizardLM & arc task, we need to

pip install sentencepiece protobuf

Mistral generate data

pip install -U transformers

hf\_gbhuybdxolAZCLfiqhCzpNytPXDafIMtNR

PPO config

PPOv2Config(

\_n\_gpu=4,

adafactor=False,

adam\_beta1=0.9,

adam\_beta2=0.999,

adam\_epsilon=1e-08,

auto\_find\_batch\_size=False,

batch\_size=None,

bf16=False,

bf16\_full\_eval=False,

cliprange=0.2,

cliprange\_value=0.2,

data\_seed=None,

dataloader\_drop\_last=False,

dataloader\_num\_workers=0,

dataloader\_persistent\_workers=False,

dataloader\_pin\_memory=True,

ddp\_backend=None,

ddp\_broadcast\_buffers=None,

ddp\_bucket\_cap\_mb=None,

ddp\_find\_unused\_parameters=None,

ddp\_timeout=1800,

debug=[],

deepspeed=None,

disable\_tqdm=False,

dispatch\_batches=None,

do\_eval=False,

do\_predict=False,

do\_train=False,

eval\_accumulation\_steps=None,

eval\_delay=0,

eval\_steps=None,

evaluation\_strategy=no,

exp\_name=ppov2\_config,

fp16=False,

fp16\_backend=auto,

fp16\_full\_eval=False,

fp16\_opt\_level=O1,

fsdp=[],

fsdp\_config={'min\_num\_params': 0, 'xla': False, 'xla\_fsdp\_grad\_ckpt': False},

fsdp\_min\_num\_params=0,

fsdp\_transformer\_layer\_cls\_to\_wrap=None,

full\_determinism=False,

gamma=1,

gradient\_accumulation\_steps=1,

gradient\_checkpointing=False,

gradient\_checkpointing\_kwargs=None,

greater\_is\_better=None,

group\_by\_length=False,

half\_precision\_backend=auto,

hub\_always\_push=False,

hub\_model\_id=None,

hub\_private\_repo=False,

hub\_strategy=every\_save,

hub\_token=<HUB\_TOKEN>,

ignore\_data\_skip=False,

include\_inputs\_for\_metrics=False,

include\_num\_input\_tokens\_seen=False,

include\_tokens\_per\_second=False,

jit\_mode\_eval=False,

kl\_coef=0.05,

label\_names=None,

label\_smoothing\_factor=0.0,

lam=0.95,

learning\_rate=5e-05,

length\_column\_name=length,

load\_best\_model\_at\_end=False,

local\_batch\_size=None,

local\_mini\_batch\_size=None,

local\_rank=0,

local\_rollout\_forward\_batch\_size=64,

log\_level=passive,

log\_level\_replica=warning,

log\_on\_each\_node=True,

logging\_dir=dasda/runs/Aug06\_20-58-59\_run291265-spin-test,

logging\_first\_step=False,

logging\_nan\_inf\_filter=True,

logging\_steps=500,

logging\_strategy=steps,

lr\_scheduler\_kwargs={},

lr\_scheduler\_type=linear,

max\_grad\_norm=1.0,

max\_steps=-1,

metric\_for\_best\_model=None,

micro\_batch\_size=None,

mini\_batch\_size=None,

mp\_parameters=,

neftune\_noise\_alpha=None,

no\_cuda=False,

non\_eos\_penalty=False,

num\_mini\_batches=1,

num\_ppo\_epochs=4,

num\_sample\_generations=10,

num\_total\_batches=None,

num\_train\_epochs=3.0,

optim=adamw\_torch,

optim\_args=None,

output\_dir=dasda,

overwrite\_output\_dir=False,

past\_index=-1,

penalty\_reward\_value=-1,

per\_device\_eval\_batch\_size=8,

per\_device\_train\_batch\_size=8,

prediction\_loss\_only=False,

push\_to\_hub=False,

push\_to\_hub\_model\_id=None,

push\_to\_hub\_organization=None,

push\_to\_hub\_token=<PUSH\_TO\_HUB\_TOKEN>,

ray\_scope=last,

remove\_unused\_columns=True,

report\_to=['tensorboard', 'wandb'],

response\_length=53,

resume\_from\_checkpoint=None,

reward\_model\_path=EleutherAI/pythia-160m,

run\_name=dasda,

sanity\_check=False,

save\_on\_each\_node=False,

save\_only\_model=False,

save\_safetensors=True,

save\_steps=500,

save\_strategy=steps,

save\_total\_limit=None,

seed=42,

sft\_model\_path=EleutherAI/pythia-160m,

skip\_memory\_metrics=True,

split\_batches=False,

stop\_token=<STOP\_TOKEN>,

stop\_token\_id=None,

temperature=0.7,

tf32=None,

torch\_compile=False,

torch\_compile\_backend=None,

torch\_compile\_mode=None,

torchdynamo=None,

total\_episodes=None,

tpu\_metrics\_debug=False,

tpu\_num\_cores=None,

use\_cpu=False,

use\_ipex=False,

use\_legacy\_prediction\_loop=False,

use\_mps\_device=False,

vf\_coef=0.1,

warmup\_ratio=0.0,

warmup\_steps=0,

weight\_decay=0.0,

whiten\_rewards=False,

world\_size=None,

)

安装trl

pip install -U git+https://github.com/huggingface/trl