DeepSpeed ZERO-2 Config

```
Which type of machine are you using?
Please select a choice using the arrow or number keys, and selecting with enter
    No distributed training
    multi-CPU
    multi-XPU
 ➤  multi-GPU
    multi-NPU
    multi-MLU
    TPU
```

```
multi-GPU
How many different machines will you use (use more than 1 for multi-node training)? [1]:
Should distributed operations be checked while running for errors? This can avoid timeout issues but will be slower. [yes/NO]:
Do you wish to optimize your script with torch dynamo?[yes/NO]:
Do you want to use DeepSpeed? [yes/NO]: yes
```

```
Do you want to specify a json file to a DeepSpeed config? [yes/NO]:
-------------------------------------------------------------------------------------------------------------------------------------------
What should be your DeepSpeed's ZeRO optimization stage?
Please select a choice using the arrow or number keys, and selecting with enter
    0
    1
 ➤  2
    3
```

```
Where to offload optimizer states?
Please select a choice using the arrow or number keys, and selecting with enter
    none
 ➤  cpu
    nvme
```

```
Where to offload parameters?
Please select a choice using the arrow or number keys, and selecting with enter
    none
 ➤  cpu
    nvme
```

```
How many gradient accumulation steps you're passing in your script? [1]: 1
```

```
Do you want to use gradient clipping? [yes/NO]:
Do you want to enable `deepspeed.zero.Init` when using ZeRO Stage-3 for constructing massive models? [yes/NO]:
```

```
Do you want to enable Mixture-of-Experts training (MoE)? [yes/NO]:
How many GPU(s) should be used for distributed training? [1]:1
```

```
Do you wish to use FP16 or BF16 (mixed precision)?
Please select a choice using the arrow or number keys, and selecting with enter
    no
    fp16
 ➤  bf16
    fp8
```

NOTE:

- Currenlty no logging, so need to redirect print output when running experiments
- Currently model loading on BF16, I will test later if model also works on full precision

Example CMD

```
CUDA_VISIBLE_DEVICES=0 accelerate launch unlearn_harm_redo_accelerate.py --
model_name meta-llama/Meta-Llama-3-8B --model_save_dir
"/SAN/intelsys/llm/yadonliu/SNLP_GCW/snlp-unlearned-
models/models/test_llama8b" --log_file
"/SAN/intelsys/llm/yadonliu/SNLP_GCW/snlp-unlearned-
models/logs/test_llama8b.lo
g" --cache_dir "/home/yadonliu/huggingface_cache" --seed 42 --
retaining_dataset rajpurkar/squad --max_bad_loss 10000 --sequential=-1 --
num_epochs=1 --batch_size=1 --seed=42 --save_every=100 --lr 2e-6
```