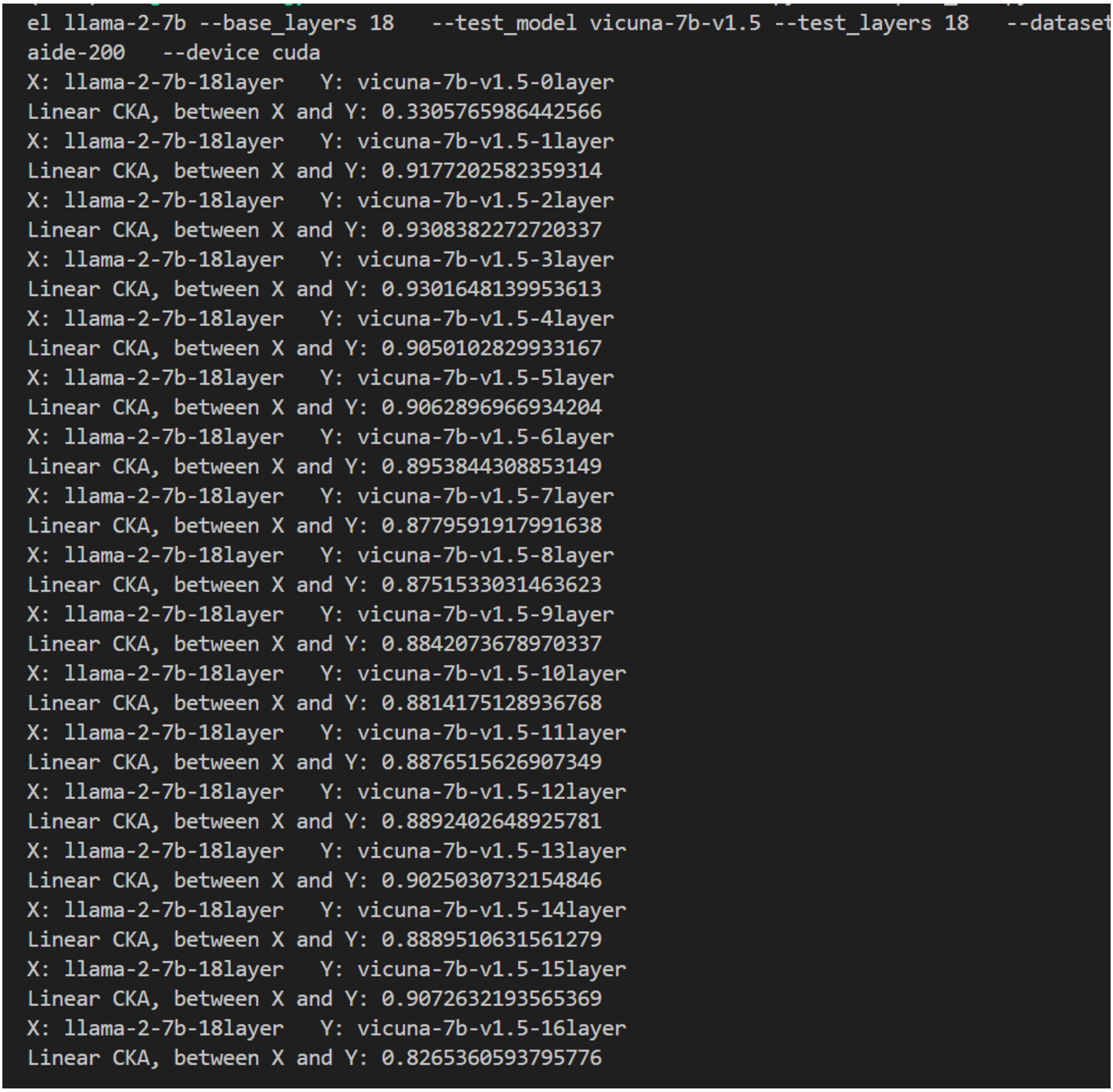
您好，我拜读了REEF这篇papar，根据您的代码和说明，成功复现了论文中的CKA相似度计算部分，但是实验数据和papar中有出入，特此反馈。

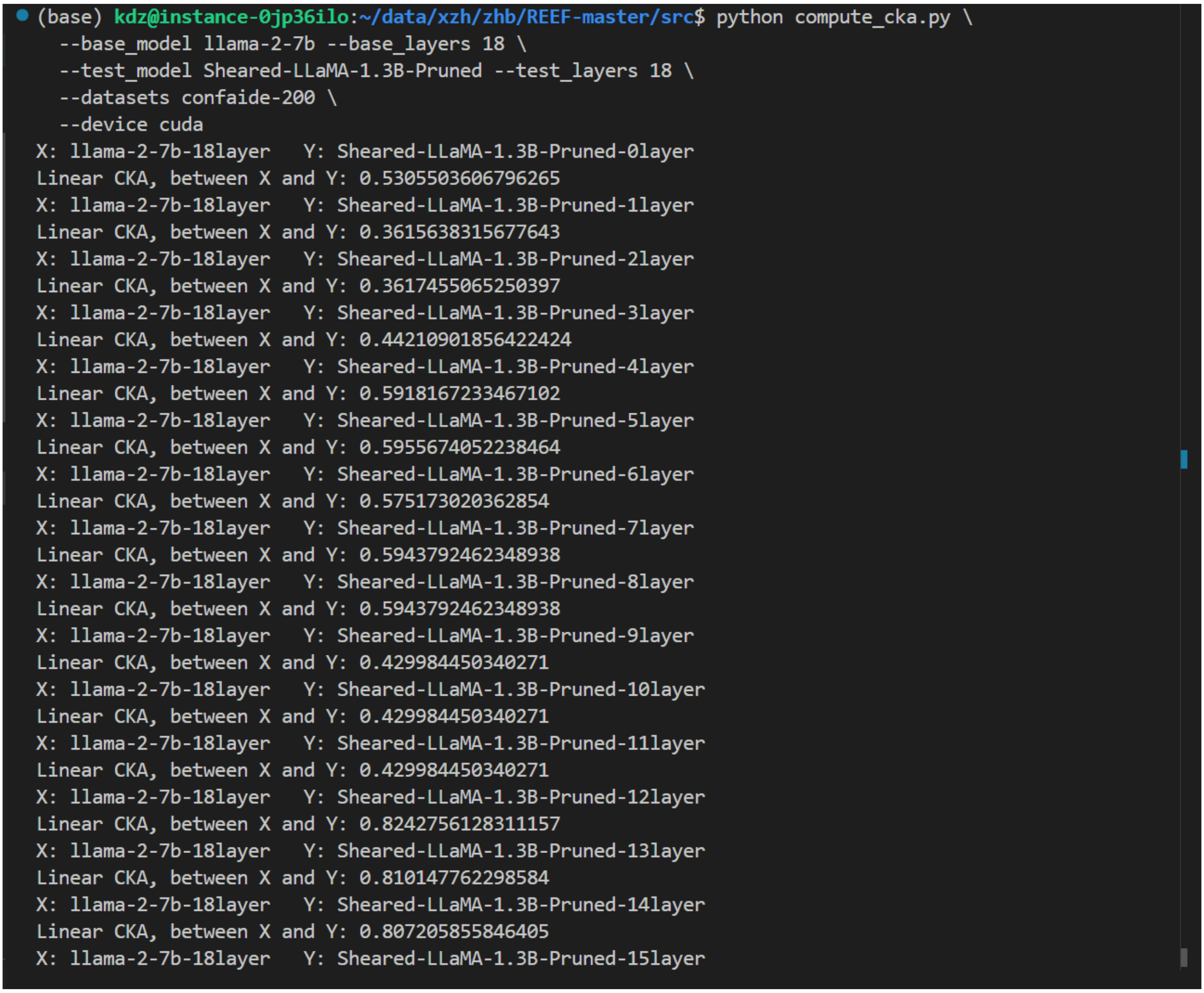
我的复现路径：
预先准备好相关模型，修改src/utils.py里的模型路径
根据 `save_activation.sh` 生成激活层
然后进入 `compute_cka.py` 进行相似度计算

结果如下

- Llama2-7b VS vicuna-7b
- 

```
el llama-2-7b --base_layers 18   --test_model vicuna-7b-v1.5 --test_layers 18   --dataset
aide-200   --device cuda
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-0layer
Linear CKA, between X and Y: 0.3305765986442566
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-1layer
Linear CKA, between X and Y: 0.9177202582359314
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-2layer
Linear CKA, between X and Y: 0.9308382272720337
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-3layer
Linear CKA, between X and Y: 0.9301648139953613
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-4layer
Linear CKA, between X and Y: 0.9050102829933167
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-5layer
Linear CKA, between X and Y: 0.9062896966934204
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-6layer
Linear CKA, between X and Y: 0.8953844308853149
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-7layer
Linear CKA, between X and Y: 0.8779591917991638
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-8layer
Linear CKA, between X and Y: 0.8751533031463623
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-9layer
Linear CKA, between X and Y: 0.8842073678970337
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-10layer
Linear CKA, between X and Y: 0.8814175128936768
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-11layer
Linear CKA, between X and Y: 0.8876515626907349
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-12layer
Linear CKA, between X and Y: 0.8892402648925781
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-13layer
Linear CKA, between X and Y: 0.9025030732154846
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-14layer
Linear CKA, between X and Y: 0.8889510631561279
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-15layer
Linear CKA, between X and Y: 0.9072632193565369
X: llama-2-7b-18layer   Y: vicuna-7b-v1.5-16layer
Linear CKA, between X and Y: 0.8265360593795776
```

- Llama2-7b VS Sheared-LLaMA-1.3B-Pruned

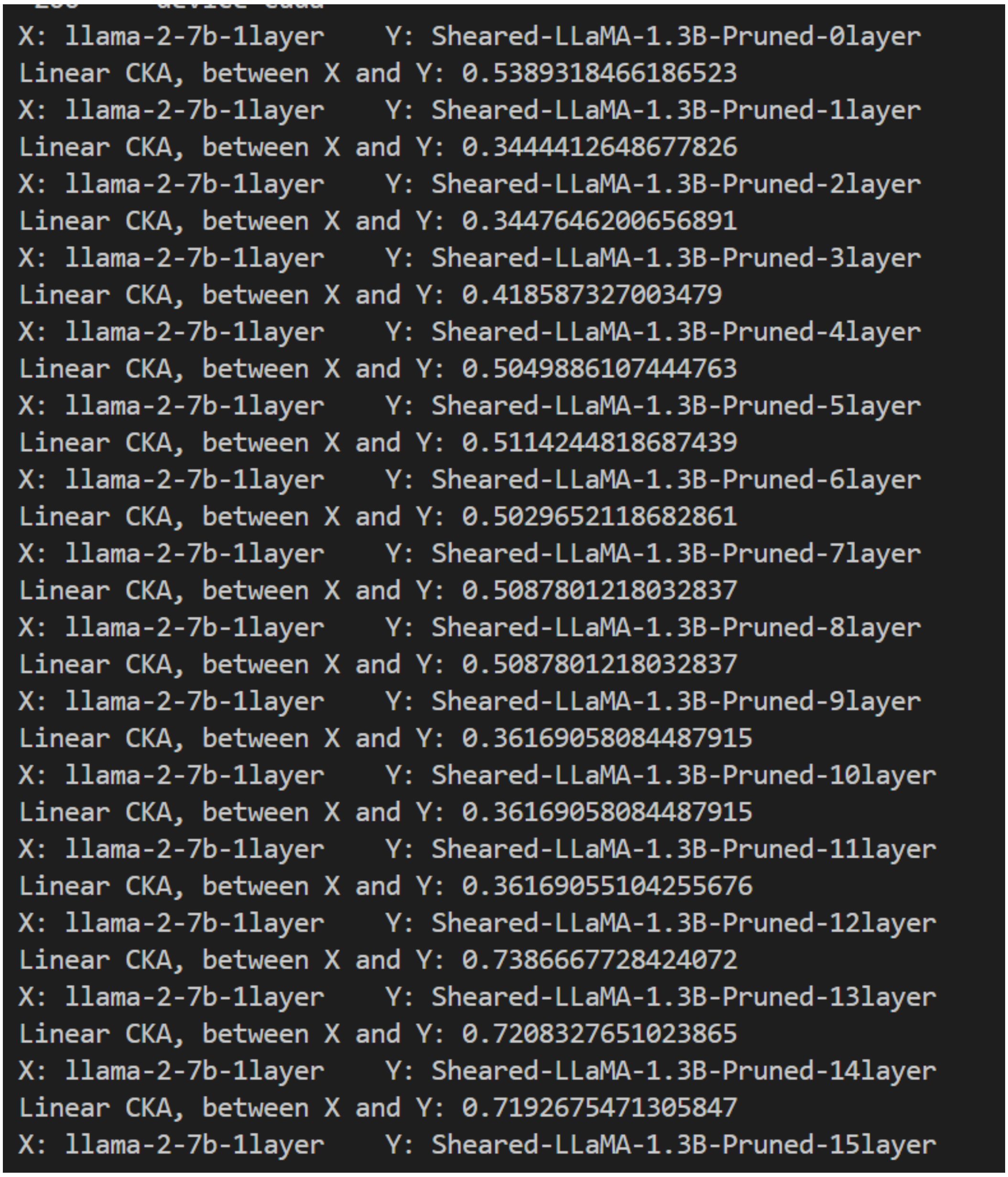


```
(base) kdz@instance-0jp36ilo:~/data/xzh/zhb/REEF-master/src$ python compute_cka.py \
    --base_model llama-2-7b --base_layers 18 \
    --test_model Sheared-LLaMA-1.3B-Pruned --test_layers 18 \
    --datasets confaide-200 \
    --device cuda
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-0layer
  Linear CKA, between X and Y: 0.5305503606796265
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-1layer
  Linear CKA, between X and Y: 0.3615638315677643
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-2layer
  Linear CKA, between X and Y: 0.3617455065250397
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-3layer
  Linear CKA, between X and Y: 0.44210901856422424
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-4layer
  Linear CKA, between X and Y: 0.5918167233467102
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-5layer
  Linear CKA, between X and Y: 0.5955674052238464
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-6layer
  Linear CKA, between X and Y: 0.575173020362854
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-7layer
  Linear CKA, between X and Y: 0.5943792462348938
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-8layer
  Linear CKA, between X and Y: 0.5943792462348938
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-9layer
  Linear CKA, between X and Y: 0.429984450340271
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-10layer
  Linear CKA, between X and Y: 0.429984450340271
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-11layer
  Linear CKA, between X and Y: 0.429984450340271
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-12layer
  Linear CKA, between X and Y: 0.8242756128311157
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-13layer
  Linear CKA, between X and Y: 0.810147762298584
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-14layer
  Linear CKA, between X and Y: 0.807205855846405
  X: llama-2-7b-18layer   Y: Sheared-LLaMA-1.3B-Pruned-15layer
```

切换为basemodel layer1

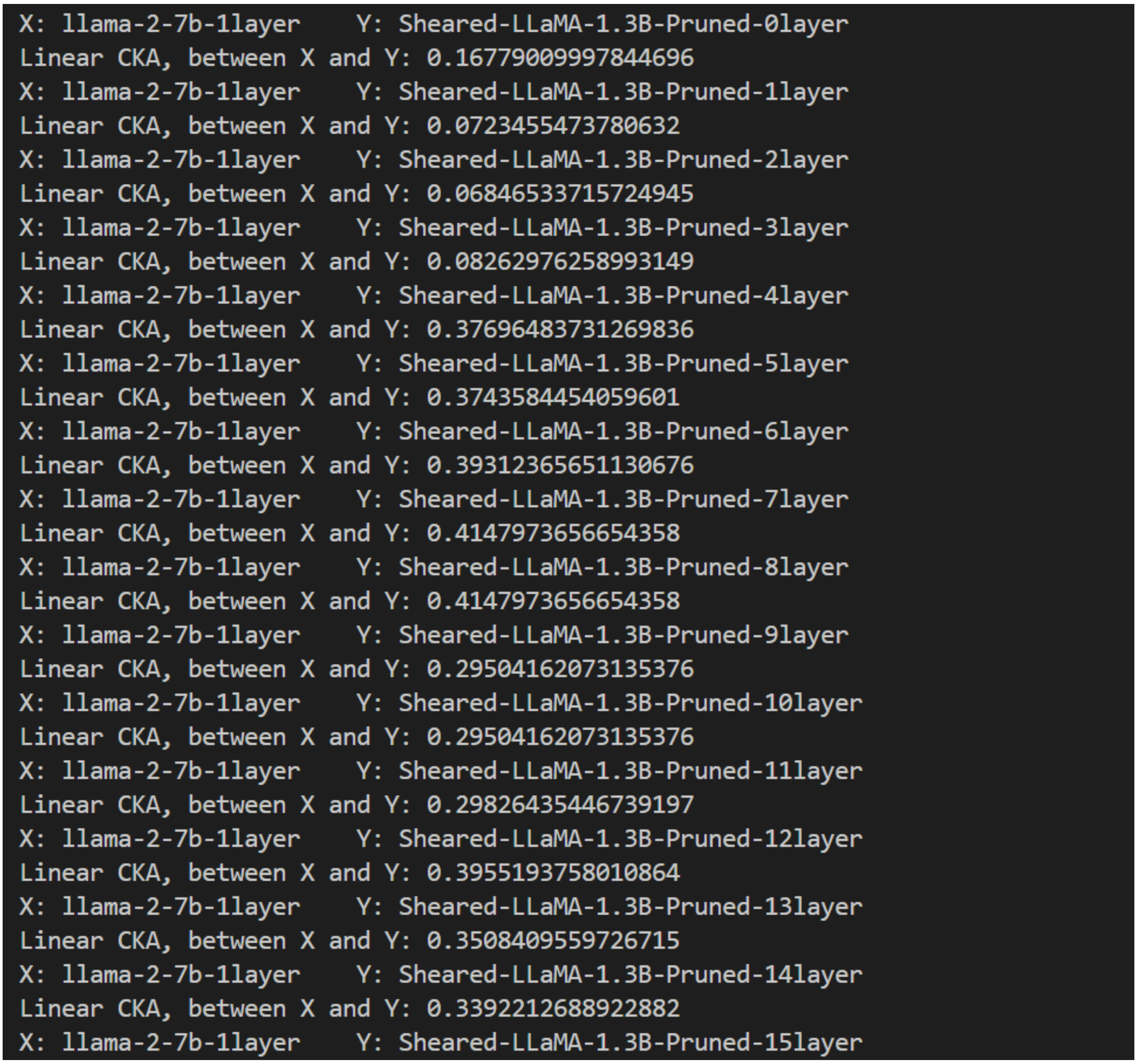
```
python compute_cka.py    --base_model llama-2-7b --base_layers 1    --test_model Sheared-LLaMA-1.3B-Pruned --test_layers 18    --datasets confaide
```

结果如下：



```
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-0layer
Linear CKA, between X and Y: 0.5389318466186523
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-1layer
Linear CKA, between X and Y: 0.3444412648677826
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-2layer
Linear CKA, between X and Y: 0.3447646200656891
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-3layer
Linear CKA, between X and Y: 0.418587327003479
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-4layer
Linear CKA, between X and Y: 0.5049886107444763
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-5layer
Linear CKA, between X and Y: 0.5114244818687439
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-6layer
Linear CKA, between X and Y: 0.5029652118682861
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-7layer
Linear CKA, between X and Y: 0.5087801218032837
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-8layer
Linear CKA, between X and Y: 0.5087801218032837
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-9layer
Linear CKA, between X and Y: 0.36169058084487915
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-10layer
Linear CKA, between X and Y: 0.36169058084487915
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-11layer
Linear CKA, between X and Y: 0.36169055104255676
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-12layer
Linear CKA, between X and Y: 0.7386667728424072
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-13layer
Linear CKA, between X and Y: 0.7208327651023865
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-14layer
Linear CKA, between X and Y: 0.7192675471305847
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-15layer
```

切换另一个数据集：toxigen



```
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-0layer
Linear CKA, between X and Y: 0.16779009997844696
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-1layer
Linear CKA, between X and Y: 0.0723455473780632
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-2layer
Linear CKA, between X and Y: 0.06846533715724945
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-3layer
Linear CKA, between X and Y: 0.08262976258993149
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-4layer
Linear CKA, between X and Y: 0.37696483731269836
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-5layer
Linear CKA, between X and Y: 0.3743584454059601
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-6layer
Linear CKA, between X and Y: 0.39312365651130676
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-7layer
Linear CKA, between X and Y: 0.4147973656654358
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-8layer
Linear CKA, between X and Y: 0.4147973656654358
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-9layer
Linear CKA, between X and Y: 0.29504162073135376
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-10layer
Linear CKA, between X and Y: 0.29504162073135376
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-11layer
Linear CKA, between X and Y: 0.29826435446739197
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-12layer
Linear CKA, between X and Y: 0.3955193758010864
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-13layer
Linear CKA, between X and Y: 0.3508409559726715
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-14layer
Linear CKA, between X and Y: 0.3392212688922882
X: llama-2-7b-1layer    Y: Sheared-LLaMA-1.3B-Pruned-15layer
```

一点疑问：

- 论文中table 1的数据是根据默认配置计算CKA相似度得到吗？是base model的layer 18和嫌疑模型的哪一层的相似度作为结果呢？
- 实验数据和论文上差距的原因是什么？