Benchmark

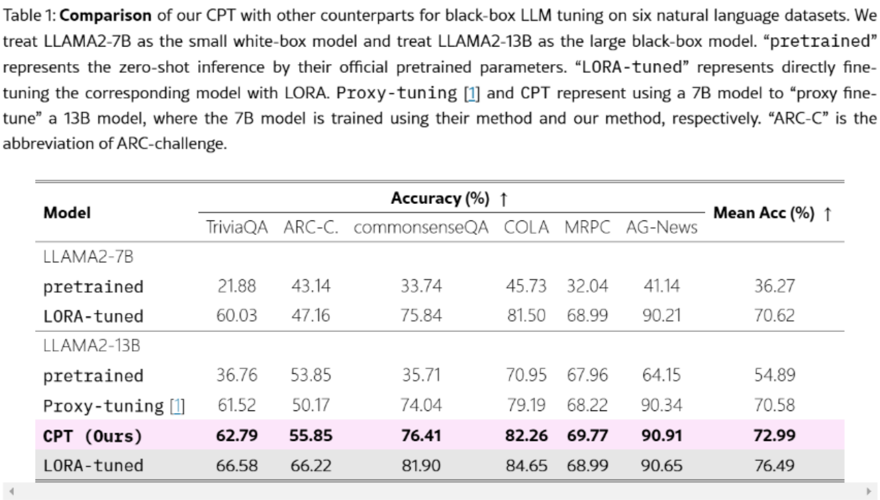
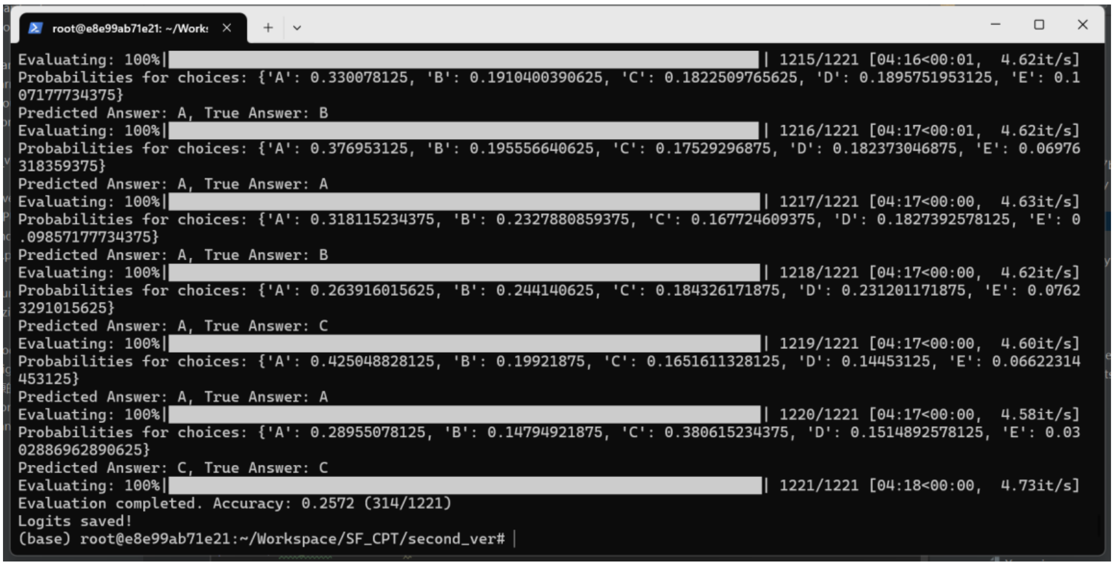


Table 1: **Comparison** of our CPT with other counterparts for black-box LLM tuning on six natural language datasets. We treat LLAMA2-7B as the small white-box model and treat LLAMA2-13B as the large black-box model. "`pretrained`" represents the zero-shot inference by their official pretrained parameters. "`LORA-tuned`" represents directly fine-tuning the corresponding model with LORA. `Proxy-tuning` [1] and CPT represent using a 7B model to "proxy fine-tune" a 13B model, where the 7B model is trained using their method and our method, respectively. "ARC-C" is the abbreviation of ARC-challenge.

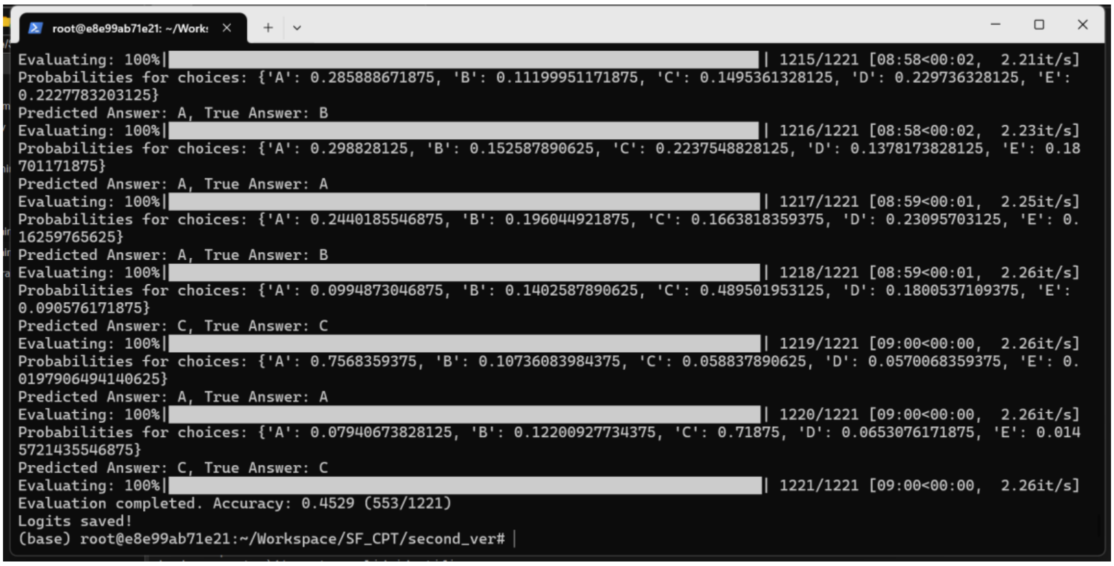
Model	**Accuracy (%) ↑**						**Mean Acc (%) ↑**
	TriviaQA	ARC-C.	commonsenseQA	COLA	MRPC	AG-News	
LLAMA2-7B							
`pretrained`	21.88	43.14	33.74	45.73	32.04	41.14	36.27
`LORA-tuned`	60.03	47.16	75.84	81.50	68.99	90.21	70.62
LLAMA2-13B							
`pretrained`	36.76	53.85	35.71	70.95	67.96	64.15	54.89
`Proxy-tuning` [1]	61.52	50.17	74.04	79.19	68.22	90.34	70.58
`CPT (Ours)`	**62.79**	**55.85**	**76.41**	**82.26**	**69.77**	**90.91**	**72.99**
`LORA-tuned`	66.58	66.22	81.90	84.65	68.99	90.65	76.49

7b pretrain accuracy: **0.2572**   7b_pretrain.py



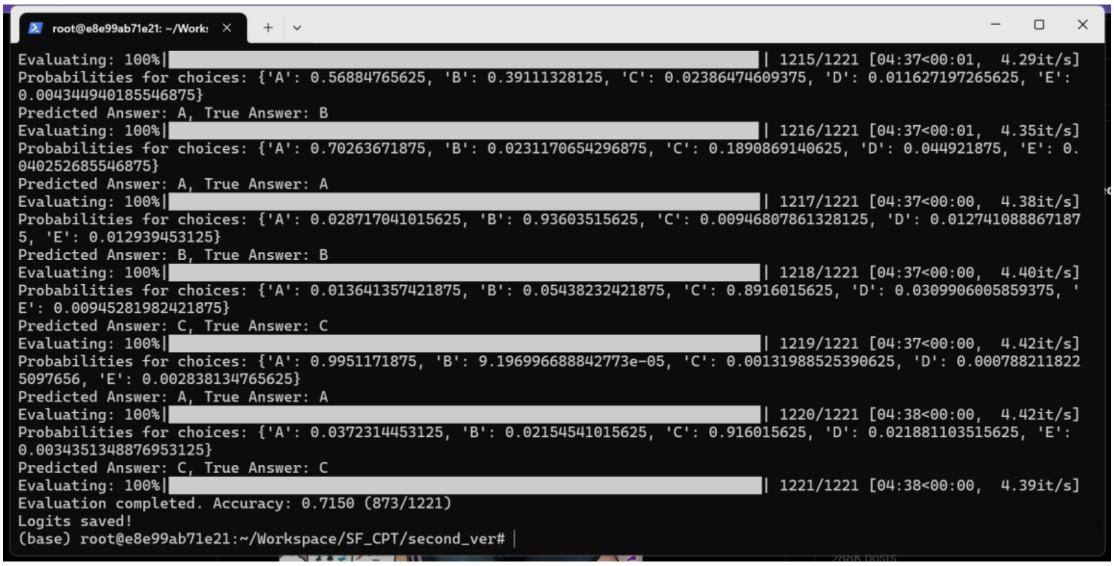
```
Evaluating: 100%|██████████████████████████████████████████████████| 1215/1221 [04:16<00:01,  4.62it/s]
Probabilities for choices: {'A': 0.330078125, 'B': 0.1910400390625, 'C': 0.1822509765625, 'D': 0.1895751953125, 'E': 0.1
07177734375}
Predicted Answer: A, True Answer: B
Evaluating: 100%|██████████████████████████████████████████████████| 1216/1221 [04:17<00:01,  4.62it/s]
Probabilities for choices: {'A': 0.376953125, 'B': 0.195556640625, 'C': 0.17529296875, 'D': 0.182373046875, 'E': 0.06976
318359375}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████████████████████████████████████████████| 1217/1221 [04:17<00:00,  4.63it/s]
Probabilities for choices: {'A': 0.318115234375, 'B': 0.2327880859375, 'C': 0.167724609375, 'D': 0.1827392578125, 'E': 0
.09857177734375}
Predicted Answer: A, True Answer: B
Evaluating: 100%|██████████████████████████████████████████████████| 1218/1221 [04:17<00:00,  4.62it/s]
Probabilities for choices: {'A': 0.263916015625, 'B': 0.244140625, 'C': 0.184326171875, 'D': 0.231201171875, 'E': 0.0762
3291015625}
Predicted Answer: A, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████████| 1219/1221 [04:17<00:00,  4.60it/s]
Probabilities for choices: {'A': 0.425048828125, 'B': 0.19921875, 'C': 0.1651611328125, 'D': 0.14453125, 'E': 0.06622314
453125}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████████████████████████████████████████████| 1220/1221 [04:17<00:00,  4.58it/s]
Probabilities for choices: {'A': 0.28955078125, 'B': 0.14794921875, 'C': 0.380615234375, 'D': 0.1514892578125, 'E': 0.03
02886962890625}
Predicted Answer: C, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████████| 1221/1221 [04:18<00:00,  4.73it/s]
Evaluation completed. Accuracy: 0.2572 (314/1221)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/second_ver# |
```

13b pretrain accuracy: **0.4529**   13b_pretrain.py



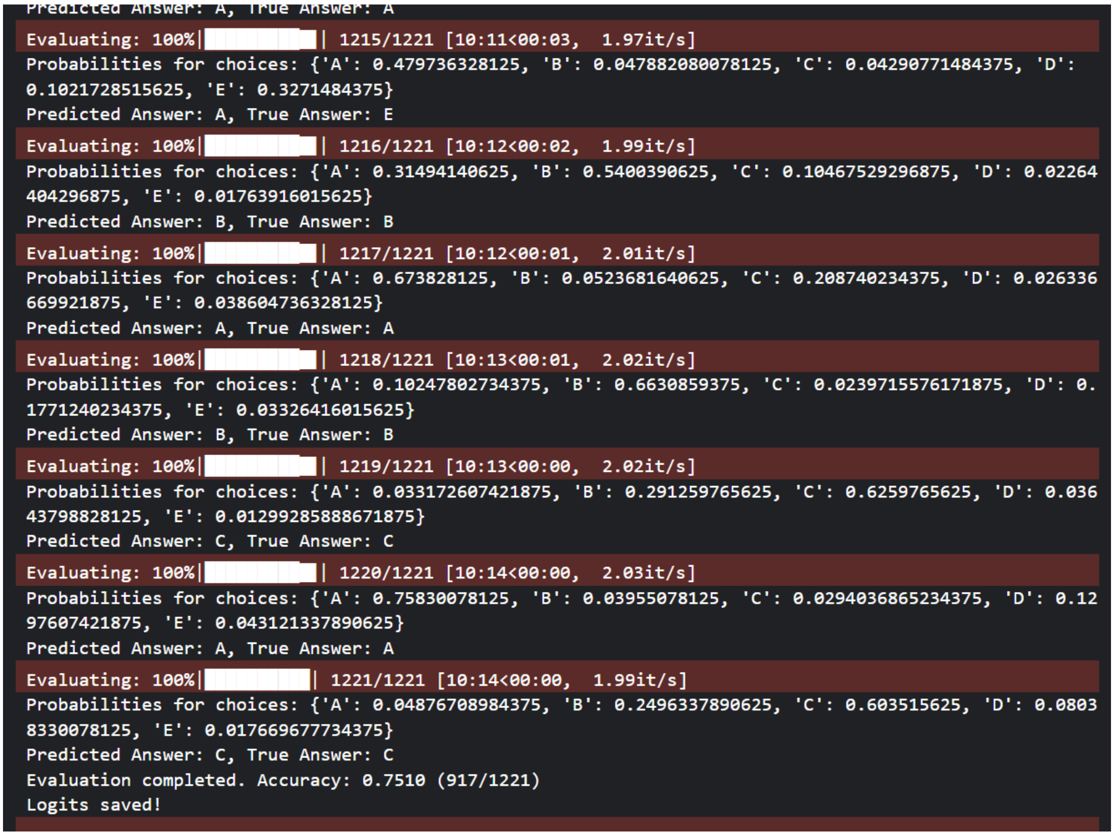
```
Evaluating: 100%|██████████████████████████████████████████████████| 1215/1221 [08:58<00:02,  2.21it/s]
Probabilities for choices: {'A': 0.285888671875, 'B': 0.11199951171875, 'C': 0.1495361328125, 'D': 0.229736328125, 'E':
0.2227783203125}
Predicted Answer: A, True Answer: B
Evaluating: 100%|██████████████████████████████████████████████████| 1216/1221 [08:58<00:02,  2.23it/s]
Probabilities for choices: {'A': 0.298828125, 'B': 0.152587890625, 'C': 0.2237548828125, 'D': 0.1378173828125, 'E': 0.18
701171875}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████████████████████████████████████████████| 1217/1221 [08:59<00:01,  2.25it/s]
Probabilities for choices: {'A': 0.2440185546875, 'B': 0.196044921875, 'C': 0.1663818359375, 'D': 0.23095703125, 'E': 0.
16259765625}
Predicted Answer: A, True Answer: B
Evaluating: 100%|██████████████████████████████████████████████████| 1218/1221 [08:59<00:01,  2.26it/s]
Probabilities for choices: {'A': 0.0994873046875, 'B': 0.1402587890625, 'C': 0.489501953125, 'D': 0.1800537109375, 'E':
0.090576171875}
Predicted Answer: C, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████████| 1219/1221 [09:00<00:00,  2.26it/s]
Probabilities for choices: {'A': 0.7568359375, 'B': 0.10736083984375, 'C': 0.058837890625, 'D': 0.0570068359375, 'E': 0.
0197906494140625}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████████████████████████████████████████████| 1220/1221 [09:00<00:00,  2.26it/s]
Probabilities for choices: {'A': 0.07940673828125, 'B': 0.12200927734375, 'C': 0.71875, 'D': 0.0653076171875, 'E': 0.014
5721435546875}
Predicted Answer: C, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████████| 1221/1221 [09:00<00:00,  2.26it/s]
Evaluation completed. Accuracy: 0.4529 (553/1221)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/second_ver# |
```

7b lora-tune accuracy: **0.7150**



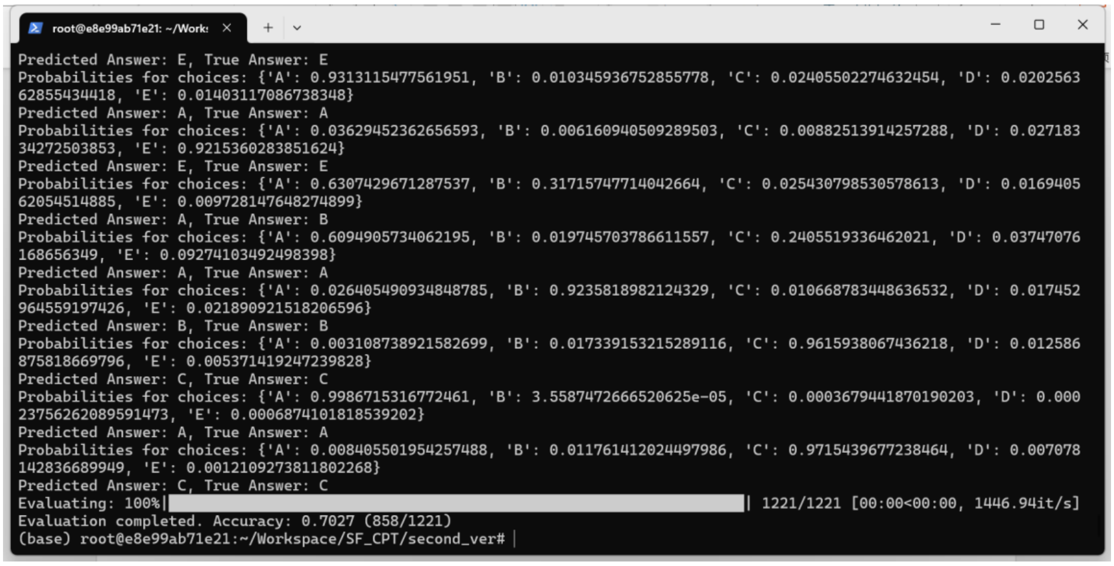
```
Evaluating: 100%|██████████████████████████████████████████████████| 1215/1221 [04:37<00:01,  4.29it/s]
Probabilities for choices: {'A': 0.56884765625, 'B': 0.39111328125, 'C': 0.02386474609375, 'D': 0.011627197265625, 'E':
0.004344940185546875}
Predicted Answer: A, True Answer: B
Evaluating: 100%|██████████████████████████████████████████████████| 1216/1221 [04:37<00:01,  4.35it/s]
Probabilities for choices: {'A': 0.70263671875, 'B': 0.0231170654296875, 'C': 0.1890869140625, 'D': 0.044921875, 'E': 0.
040252685546875}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████████████████████████████████████████████| 1217/1221 [04:37<00:00,  4.38it/s]
Probabilities for choices: {'A': 0.028717041015625, 'B': 0.93603515625, 'C': 0.00946807861328125, 'D': 0.012741088867187
5, 'E': 0.012939453125}
Predicted Answer: B, True Answer: B
Evaluating: 100%|██████████████████████████████████████████████████| 1218/1221 [04:37<00:00,  4.40it/s]
Probabilities for choices: {'A': 0.013641357421875, 'B': 0.05438232421875, 'C': 0.8916015625, 'D': 0.0309906005859375, '
E': 0.00945281982421875}
Predicted Answer: C, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████████| 1219/1221 [04:37<00:00,  4.42it/s]
Probabilities for choices: {'A': 0.9951171875, 'B': 9.196996688842773e-05, 'C': 0.00131988525390625, 'D': 0.000788211822
5097656, 'E': 0.002838134765625}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████████████████████████████████████████████| 1220/1221 [04:38<00:00,  4.42it/s]
Probabilities for choices: {'A': 0.0372314453125, 'B': 0.021545410015625, 'C': 0.916015625, 'D': 0.021881103515625, 'E':
0.0034351348876953125}
Predicted Answer: C, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████████| 1221/1221 [04:38<00:00,  4.39it/s]
Evaluation completed. Accuracy: 0.7150 (873/1221)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/second_ver#
```

13b lora-tune accuracy: **0.7510**



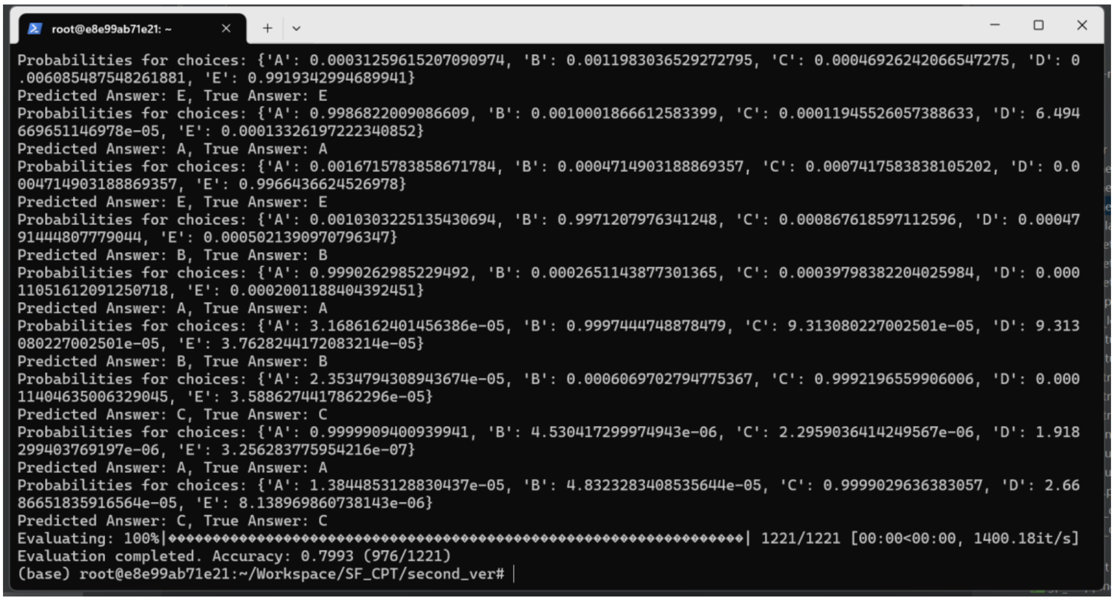
```
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████| 1215/1221 [10:11<00:03,  1.97it/s]
Probabilities for choices: {'A': 0.479736328125, 'B': 0.047882080078125, 'C': 0.04290771484375, 'D':
0.1021728515625, 'E': 0.3271484375}
Predicted Answer: A, True Answer: E
Evaluating: 100%|██████████| 1216/1221 [10:12<00:02,  1.99it/s]
Probabilities for choices: {'A': 0.31494140625, 'B': 0.5400390625, 'C': 0.10467529296875, 'D': 0.02264
404296875, 'E': 0.01763916015625}
Predicted Answer: B, True Answer: B
Evaluating: 100%|██████████| 1217/1221 [10:12<00:01,  2.01it/s]
Probabilities for choices: {'A': 0.673828125, 'B': 0.0523681640625, 'C': 0.208740234375, 'D': 0.026336
669921875, 'E': 0.038604736328125}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████| 1218/1221 [10:13<00:01,  2.02it/s]
Probabilities for choices: {'A': 0.10247802734375, 'B': 0.6630859375, 'C': 0.0239715576171875, 'D': 0.
1771240234375, 'E': 0.03326416015625}
Predicted Answer: B, True Answer: B
Evaluating: 100%|██████████| 1219/1221 [10:13<00:00,  2.02it/s]
Probabilities for choices: {'A': 0.033172607421875, 'B': 0.291259765625, 'C': 0.6259765625, 'D': 0.036
43798828125, 'E': 0.0129928588867187 5}
Predicted Answer: C, True Answer: C
Evaluating: 100%|██████████| 1220/1221 [10:14<00:00,  2.03it/s]
Probabilities for choices: {'A': 0.75830078125, 'B': 0.03955078125, 'C': 0.0294036865234375, 'D': 0.12
97607421875, 'E': 0.043121337890625}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████| 1221/1221 [10:14<00:00,  1.99it/s]
Probabilities for choices: {'A': 0.04876708984375, 'B': 0.2496337890625, 'C': 0.603515625, 'D': 0.0803
8330078125, 'E': 0.017669677734375}
Predicted Answer: C, True Answer: C
Evaluation completed. Accuracy: 0.7510 (917/1221)
Logits saved!
```

Proxy tuning accuracy: **0.7027**   proxy_tuning.py



```
Predicted Answer: E, True Answer: E
Probabilities for choices: {'A': 0.9313115477561951, 'B': 0.010345936752855778, 'C': 0.02405502274632454, 'D': 0.0202563
62855434418, 'E': 0.014031170867383348}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 0.03629452362656593, 'B': 0.006160940509289503, 'C': 0.00882513914257288, 'D': 0.027183
34272503853, 'E': 0.9215360283851624}
Predicted Answer: E, True Answer: E
Probabilities for choices: {'A': 0.6307429671287537, 'B': 0.31715747714042664, 'C': 0.025430798530578613, 'D': 0.0169405
62054514885, 'E': 0.009728147648274899}
Predicted Answer: A, True Answer: B
Probabilities for choices: {'A': 0.6094905734062195, 'B': 0.019745703786611557, 'C': 0.2405519336462021, 'D': 0.03747076
168656349, 'E': 0.09274103492498398}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 0.026405490934848785, 'B': 0.9235818982124329, 'C': 0.010668783448636532, 'D': 0.017452
964559197426, 'E': 0.021890921518206596}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 0.003108738921582699, 'B': 0.017339153215289116, 'C': 0.9615938067436218, 'D': 0.012586
875818669796, 'E': 0.005371419247239828}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 0.9986715316772461, 'B': 3.5587472666520625e-05, 'C': 0.0003679441870190203, 'D': 0.000
23756262089591473, 'E': 0.0006874101818539202}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 0.008405501954257488, 'B': 0.011761412024497986, 'C': 0.9715439677238464, 'D': 0.007078
142836689949, 'E': 0.0012109273811802268}
Predicted Answer: C, True Answer: C
Evaluating: 100%|█████████████████████████████████████████████████████████████████| 1221/1221 [00:00<00:00, 1446.94it/s]
Evaluation completed. Accuracy: 0.7027 (858/1221)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/second_ver# |
```

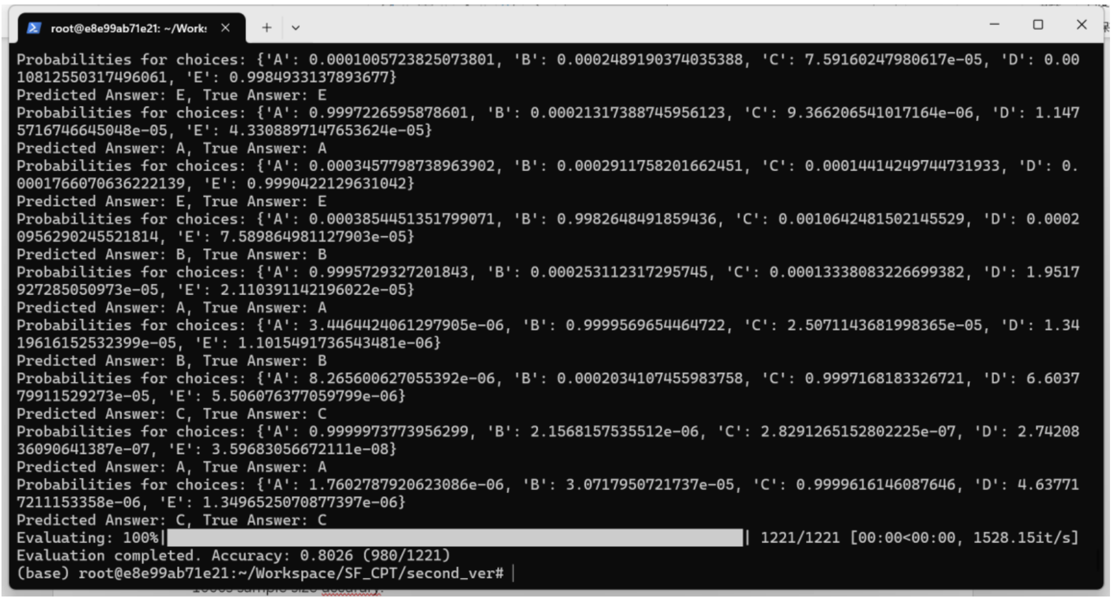
CPT tuning accuracy: **0.7993**



```
Probabilities for choices: {'A': 0.00031259615207090974, 'B': 0.0011983036529272795, 'C': 0.00046926242066547275, 'D': 0
.006085487548261881, 'E': 0.9919342994689941}
Predicted Answer: E, True Answer: E
Probabilities for choices: {'A': 0.9986822009086609, 'B': 0.0010001866612583399, 'C': 0.00011945526057388633, 'D': 6.494
669651146978e-05, 'E': 0.00013326197222340852}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 0.0016715783858671784, 'B': 0.0004714903188869357, 'C': 0.0007417583838105202, 'D': 0.0
004714903188869357, 'E': 0.9966436624526978}
Predicted Answer: E, True Answer: E
Probabilities for choices: {'A': 0.0010303225135430694, 'B': 0.9971207976341248, 'C': 0.000867618597112596, 'D': 0.00047
91444807779044, 'E': 0.0005021390970796347}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 0.9990262985229492, 'B': 0.0002651143877301365, 'C': 0.00039798382204025984, 'D': 0.000
11051612091250718, 'E': 0.0002001188404392451}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 3.1686162401456386e-05, 'B': 0.9997444748878479, 'C': 9.313080227002501e-05, 'D': 9.313
080227002501e-05, 'E': 3.7628244172083214e-05}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 2.3534794308943674e-05, 'B': 0.0006069702794775367, 'C': 0.9992196559906006, 'D': 0.000
11404635006329045, 'E': 3.5886274417862296e-05}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 0.9999909400939941, 'B': 4.530417299974943e-06, 'C': 2.2959036414249567e-06, 'D': 1.918
299403769197e-06, 'E': 3.256283775954216e-07}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 1.3844853128830437e-05, 'B': 4.8323283408535644e-05, 'C': 0.9999029636383057, 'D': 2.66
86651835916564e-05, 'E': 8.138969860738143e-06}
Predicted Answer: C, True Answer: C
Evaluating: 100%|�����������������������������������������������������������������| 1221/1221 [00:00<00:00, 1400.18it/s]
Evaluation completed. Accuracy: 0.7993 (976/1221)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/second_ver# |
```

Gaussian CPT tuning accuracy: **0.8026**

1000 / 9741 (10.27%)

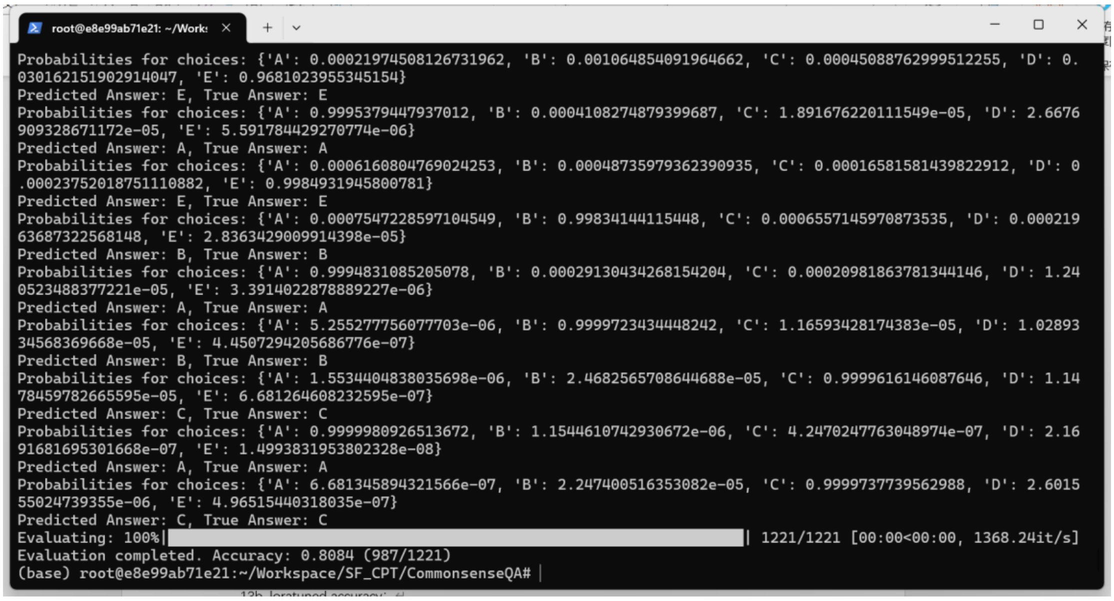


```
Probabilities for choices: {'A': 0.0001005723825073801, 'B': 0.0002489190374035388, 'C': 7.59160247980617e-05, 'D': 0.00
10812550317496061, 'E': 0.9984933137893677}
Predicted Answer: E, True Answer: E
Probabilities for choices: {'A': 0.9997226595878601, 'B': 0.00021317388745956123, 'C': 9.366206541017164e-06, 'D': 1.147
5716746645048e-05, 'E': 4.3308897147653624e-05}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 0.0003457798738963902, 'B': 0.0002911758201662451, 'C': 0.0001441424974731933, 'D': 0.
0001766070636222139, 'E': 0.9990422129631042}
Predicted Answer: E, True Answer: E
Probabilities for choices: {'A': 0.0003854451351799071, 'B': 0.9982648491859436, 'C': 0.0010642481502145529, 'D': 0.0002
0956290245521814, 'E': 7.589864981127903e-05}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 0.9995729327201843, 'B': 0.000253112317295745, 'C': 0.0001333808322669382, 'D': 1.9517
927285050973e-05, 'E': 2.110391142196022e-05}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 3.4464424061297905e-06, 'B': 0.9999569654464722, 'C': 2.5071143681998365e-05, 'D': 1.34
19616152532399e-05, 'E': 1.1015491736543481e-06}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 8.265600627055392e-06, 'B': 0.0002034107455983758, 'C': 0.9997168183326721, 'D': 6.6037
79911529273e-05, 'E': 5.506076377059799e-06}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 0.9999973773956299, 'B': 2.1568157535512e-06, 'C': 2.8291265152802225e-07, 'D': 2.74208
36090641387e-07, 'E': 3.59683056672111e-08}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 1.7602787920623086e-06, 'B': 3.0717950721737e-05, 'C': 0.9999616146087646, 'D': 4.63771
7211153358e-06, 'E': 1.3496525070877397e-06}
Predicted Answer: C, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████| 1221/1221 [00:00<00:00, 1528.15it/s]
Evaluation completed. Accuracy: 0.8026 (980/1221)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/second_ver# |
```

GP with filter accuracy: **0.8084**

Config: input_threshold=0.14, output_threshold=3

201 / 9741 (2.1%)



```
Probabilities for choices: {'A': 0.00021974508126731962, 'B': 0.001064854091964662, 'C': 0.00045088762999512255, 'D': 0.
030162151902914047, 'E': 0.9681023955345154}
Predicted Answer: E, True Answer: E
Probabilities for choices: {'A': 0.9995379447937012, 'B': 0.0004108274879399687, 'C': 1.891676220111549e-05, 'D': 2.6676
909328671172e-05, 'E': 5.591784429270774e-06}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 0.0006160804769024253, 'B': 0.00048735979362390935, 'C': 0.00016581581439822912, 'D': 0
.00023752018751110882, 'E': 0.9984931945800781}
Predicted Answer: E, True Answer: E
Probabilities for choices: {'A': 0.0007547228597104549, 'B': 0.99834144115448, 'C': 0.0006557145970873535, 'D': 0.000219
63687322568148, 'E': 2.8363429009914398e-05}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 0.9994831085205078, 'B': 0.00029130434268154204, 'C': 0.00020981863781344146, 'D': 1.24
0523488377221e-05, 'E': 3.3914022878889227e-06}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 5.255277756077703e-06, 'B': 0.9999723434448242, 'C': 1.16593428174383e-05, 'D': 1.02893
34568369668e-05, 'E': 4.4507294205686776e-07}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 1.5534404838035698e-06, 'B': 2.4682565708644688e-05, 'C': 0.9999616146087646, 'D': 1.14
78459782665595e-05, 'E': 6.681264608232595e-07}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 0.9999980926513672, 'B': 1.1544610742930672e-06, 'C': 4.2470247763048974e-07, 'D': 2.16
91681695301668e-07, 'E': 1.4993831953802328e-08}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 6.681345894321566e-07, 'B': 2.247400516353082e-05, 'C': 0.9999737739562988, 'D': 2.6015
55024739355e-06, 'E': 4.96515440318035e-07}
Predicted Answer: C, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████| 1221/1221 [00:00<00:00, 1368.24it/s]
Evaluation completed. Accuracy: 0.8084 (987/1221)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CommonsenseQA# |
```

13b-loratuned accuracy: