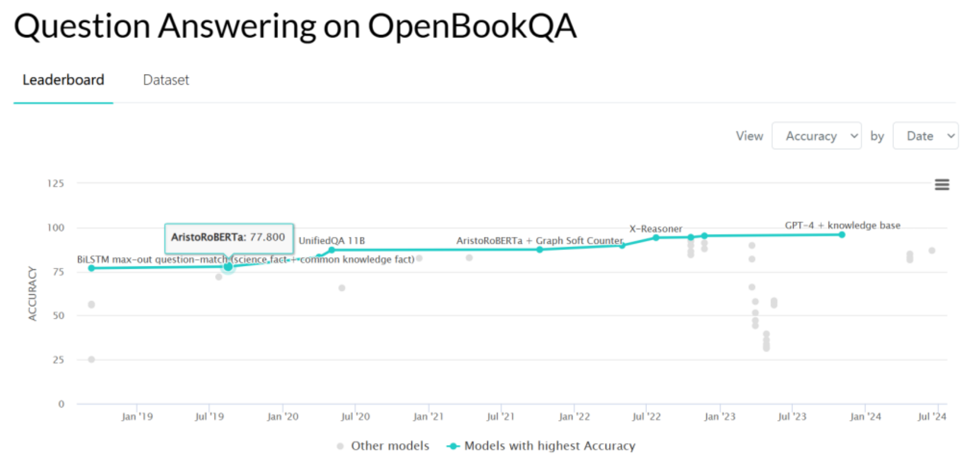
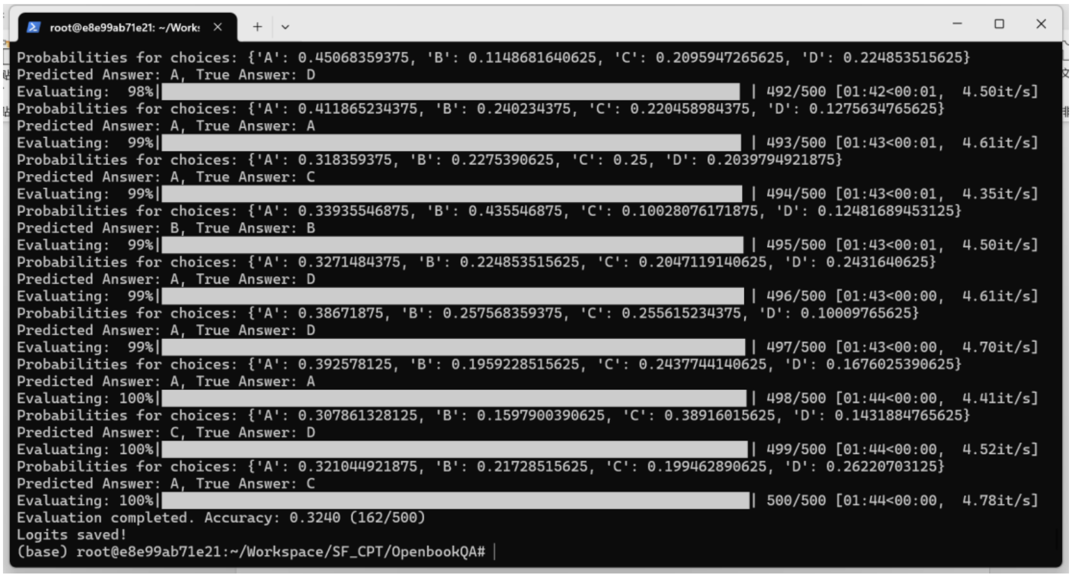
Benchmark



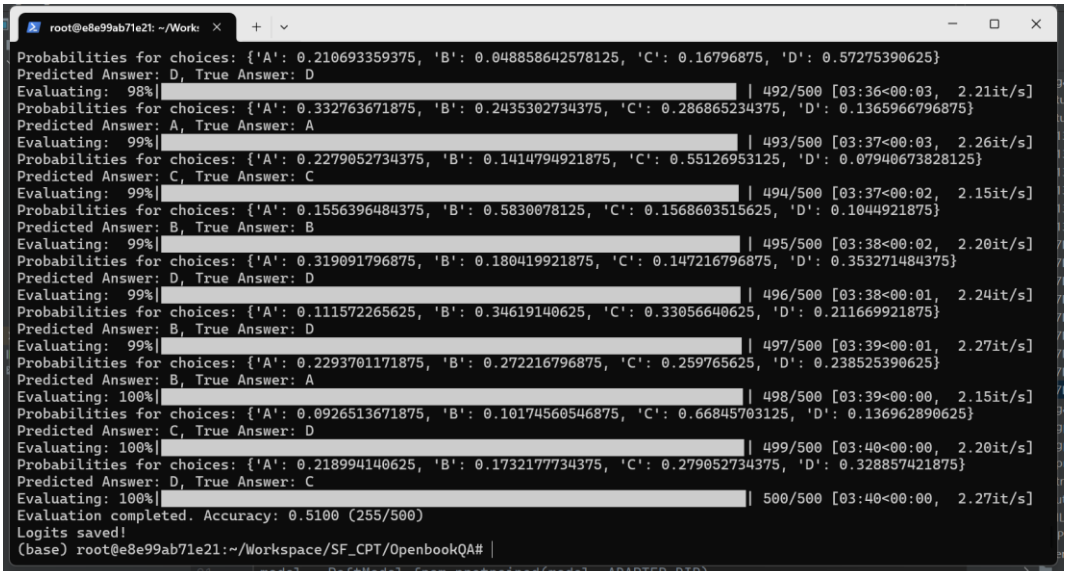
Question Answering on OpenBookQA

7b pretrain acc: **0.3240**



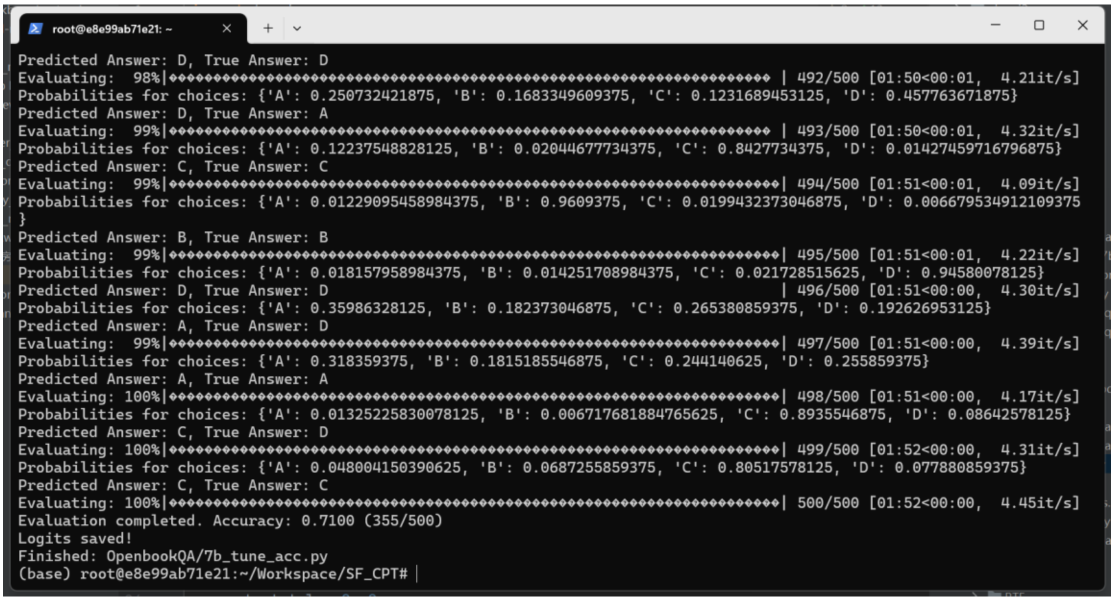
```
Probabilities for choices: {'A': 0.45068359375, 'B': 0.1148681640625, 'C': 0.2095947265625, 'D': 0.224853515625}
Predicted Answer: A, True Answer: D
Evaluating:  98%|██████████████████████████████████████████████████ | 492/500 [01:42<00:01,  4.50it/s]
Probabilities for choices: {'A': 0.411865234375, 'B': 0.240234375, 'C': 0.220458984375, 'D': 0.1275634765625}
Predicted Answer: A, True Answer: A
Evaluating:  99%|██████████████████████████████████████████████████ | 493/500 [01:43<00:01,  4.61it/s]
Probabilities for choices: {'A': 0.318359375, 'B': 0.2275390625, 'C': 0.25, 'D': 0.2039794921875}
Predicted Answer: A, True Answer: C
Evaluating:  99%|██████████████████████████████████████████████████ | 494/500 [01:43<00:01,  4.35it/s]
Probabilities for choices: {'A': 0.33935546875, 'B': 0.435546875, 'C': 0.10028076171875, 'D': 0.12481689453125}
Predicted Answer: B, True Answer: B
Evaluating:  99%|██████████████████████████████████████████████████ | 495/500 [01:43<00:01,  4.50it/s]
Probabilities for choices: {'A': 0.3271484375, 'B': 0.224853515625, 'C': 0.2047119140625, 'D': 0.2431640625}
Predicted Answer: A, True Answer: D
Evaluating:  99%|██████████████████████████████████████████████████ | 496/500 [01:43<00:00,  4.61it/s]
Probabilities for choices: {'A': 0.38671875, 'B': 0.257568359375, 'C': 0.255615234375, 'D': 0.10009765625}
Predicted Answer: A, True Answer: D
Evaluating:  99%|██████████████████████████████████████████████████ | 497/500 [01:43<00:00,  4.70it/s]
Probabilities for choices: {'A': 0.392578125, 'B': 0.1959228515625, 'C': 0.2437744140625, 'D': 0.1676025390625}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████████████████████████████████████████████ | 498/500 [01:44<00:00,  4.41it/s]
Probabilities for choices: {'A': 0.307861328125, 'B': 0.1597900390625, 'C': 0.38916015625, 'D': 0.1431884765625}
Predicted Answer: C, True Answer: D
Evaluating: 100%|██████████████████████████████████████████████████ | 499/500 [01:44<00:00,  4.52it/s]
Probabilities for choices: {'A': 0.321044921875, 'B': 0.21728515625, 'C': 0.199462890625, 'D': 0.26220703125}
Predicted Answer: A, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████████| 500/500 [01:44<00:00,  4.78it/s]
Evaluation completed. Accuracy: 0.3240 (162/500)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/OpenbookQA# 
```

13b pretrain acc: **0.5100**



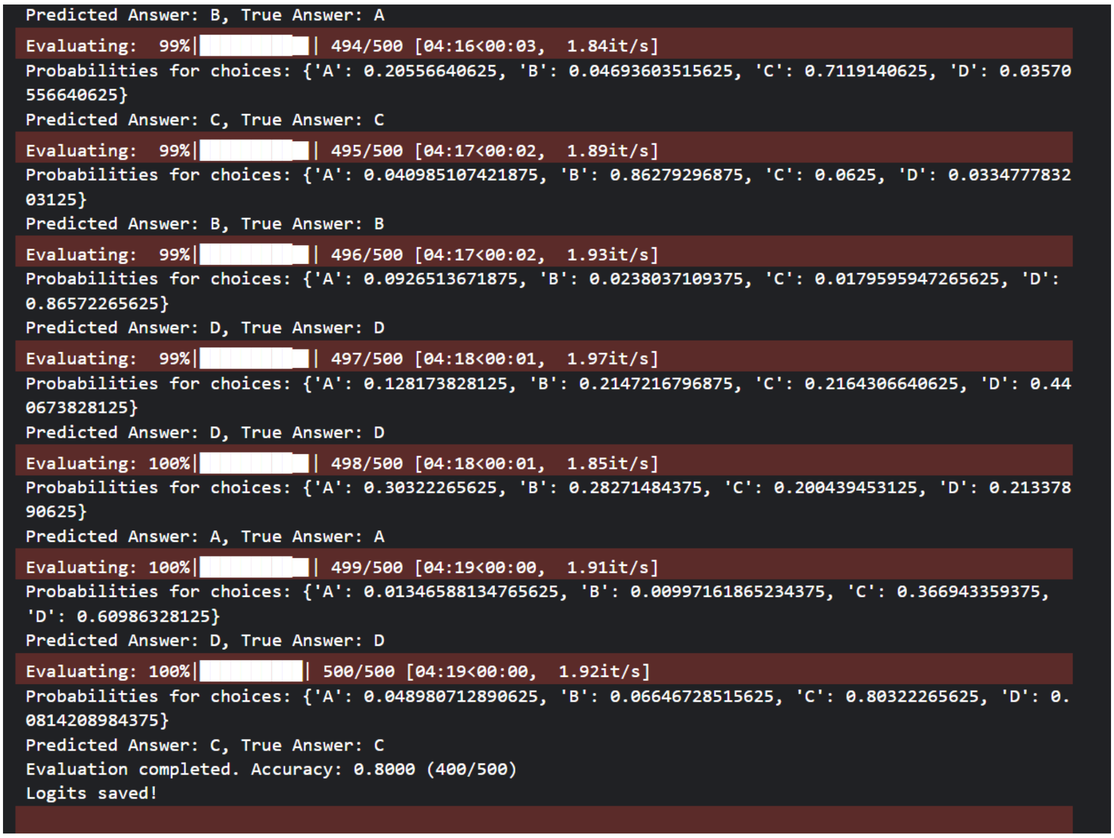
```
Probabilities for choices: {'A': 0.210693359375, 'B': 0.048858642578125, 'C': 0.16796875, 'D': 0.57275390625}
Predicted Answer: D, True Answer: D
Evaluating:  98%|██████████████████████████████████████████████████ | 492/500 [03:36<00:03,  2.21it/s]
Probabilities for choices: {'A': 0.332763671875, 'B': 0.2435302734375, 'C': 0.286865234375, 'D': 0.1365966796875}
Predicted Answer: A, True Answer: A
Evaluating:  99%|██████████████████████████████████████████████████ | 493/500 [03:37<00:03,  2.26it/s]
Probabilities for choices: {'A': 0.2279052734375, 'B': 0.1414794921875, 'C': 0.55126953125, 'D': 0.07940673828125}
Predicted Answer: C, True Answer: C
Evaluating:  99%|██████████████████████████████████████████████████ | 494/500 [03:37<00:02,  2.15it/s]
Probabilities for choices: {'A': 0.1556396484375, 'B': 0.5830078125, 'C': 0.1568603515625, 'D': 0.1044921875}
Predicted Answer: B, True Answer: B
Evaluating:  99%|██████████████████████████████████████████████████ | 495/500 [03:38<00:02,  2.20it/s]
Probabilities for choices: {'A': 0.319091796875, 'B': 0.180419921875, 'C': 0.147216796875, 'D': 0.353271484375}
Predicted Answer: D, True Answer: D
Evaluating:  99%|██████████████████████████████████████████████████ | 496/500 [03:38<00:01,  2.24it/s]
Probabilities for choices: {'A': 0.111572265625, 'B': 0.34619140625, 'C': 0.33056640625, 'D': 0.211669921875}
Predicted Answer: B, True Answer: D
Evaluating:  99%|██████████████████████████████████████████████████ | 497/500 [03:39<00:01,  2.27it/s]
Probabilities for choices: {'A': 0.2293701171875, 'B': 0.272216796875, 'C': 0.259765625, 'D': 0.238525390625}
Predicted Answer: B, True Answer: A
Evaluating: 100%|██████████████████████████████████████████████████ | 498/500 [03:39<00:00,  2.15it/s]
Probabilities for choices: {'A': 0.0926513671875, 'B': 0.10174560546875, 'C': 0.66845703125, 'D': 0.136962890625}
Predicted Answer: C, True Answer: D
Evaluating: 100%|██████████████████████████████████████████████████ | 499/500 [03:40<00:00,  2.20it/s]
Probabilities for choices: {'A': 0.218994140625, 'B': 0.1732177734375, 'C': 0.279052734375, 'D': 0.328857421875}
Predicted Answer: D, True Answer: C
Evaluating: 100%|██████████████████████████████████████████████████| 500/500 [03:40<00:00,  2.27it/s]
Evaluation completed. Accuracy: 0.5100 (255/500)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/OpenbookQA# 
```

7b lora-tune acc: **0.7100**



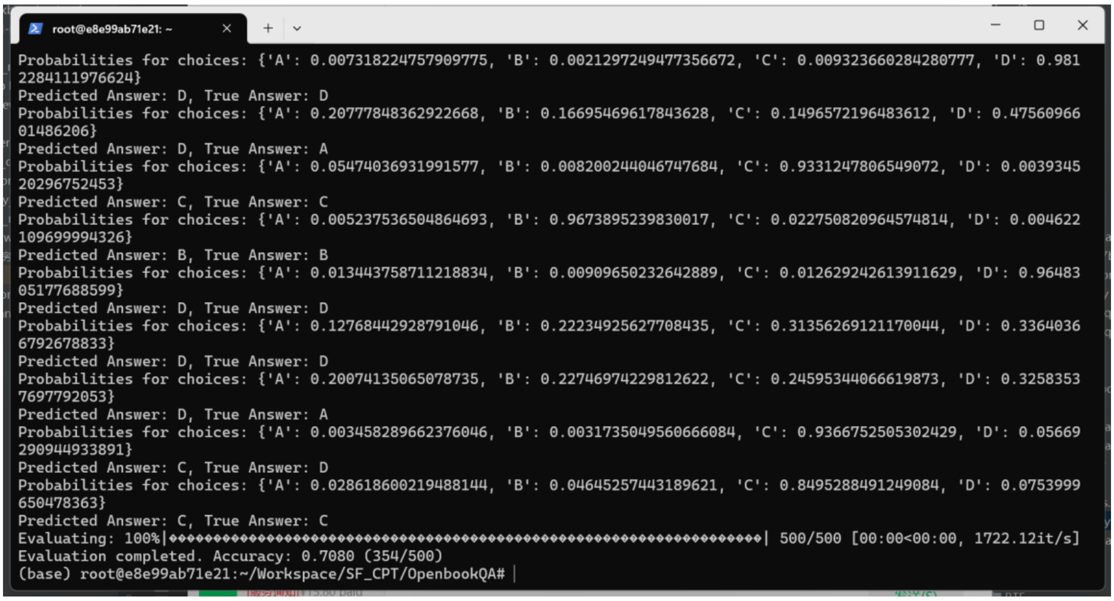
```
Predicted Answer: D, True Answer: D
Evaluating:  98%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆ | 492/500 [01:50<00:01,  4.21it/s]
Probabilities for choices: {'A': 0.250732421875, 'B': 0.1683349609375, 'C': 0.1231689453125, 'D': 0.457763671875}
Predicted Answer: D, True Answer: A
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆ | 493/500 [01:50<00:01,  4.32it/s]
Probabilities for choices: {'A': 0.12237548828125, 'B': 0.02044677734375, 'C': 0.8427734375, 'D': 0.01427459716796875}
Predicted Answer: C, True Answer: C
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 494/500 [01:51<00:01,  4.09it/s]
Probabilities for choices: {'A': 0.01229095458984375, 'B': 0.9609375, 'C': 0.0199432373046875, 'D': 0.006679534912109375
}
Predicted Answer: B, True Answer: B
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 495/500 [01:51<00:01,  4.22it/s]
Probabilities for choices: {'A': 0.018157958984375, 'B': 0.014251708984375, 'C': 0.021728515625, 'D': 0.94580078125}
Predicted Answer: D, True Answer: D                                                     | 496/500 [01:51<00:00,  4.30it/s]
Probabilities for choices: {'A': 0.35986328125, 'B': 0.182373046875, 'C': 0.265380859375, 'D': 0.192626953125}
Predicted Answer: A, True Answer: D
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 497/500 [01:51<00:00,  4.39it/s]
Probabilities for choices: {'A': 0.318359375, 'B': 0.1815185546875, 'C': 0.244140625, 'D': 0.255859375}
Predicted Answer: A, True Answer: A
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 498/500 [01:51<00:00,  4.17it/s]
Probabilities for choices: {'A': 0.01325225830078125, 'B': 0.006717681884765625, 'C': 0.8935546875, 'D': 0.08642578125}
Predicted Answer: C, True Answer: D
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 499/500 [01:52<00:00,  4.31it/s]
Probabilities for choices: {'A': 0.048004150390625, 'B': 0.0687255859375, 'C': 0.80517578125, 'D': 0.077880859375}
Predicted Answer: C, True Answer: C
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 500/500 [01:52<00:00,  4.45it/s]
Evaluation completed. Accuracy: 0.7100 (355/500)
Logits saved!
Finished: OpenbookQA/7b_tune_acc.py
(base) root@e8e99ab71e21:~/Workspace/SF_CPT# |
```

13b lora-tune acc: **0.8000**



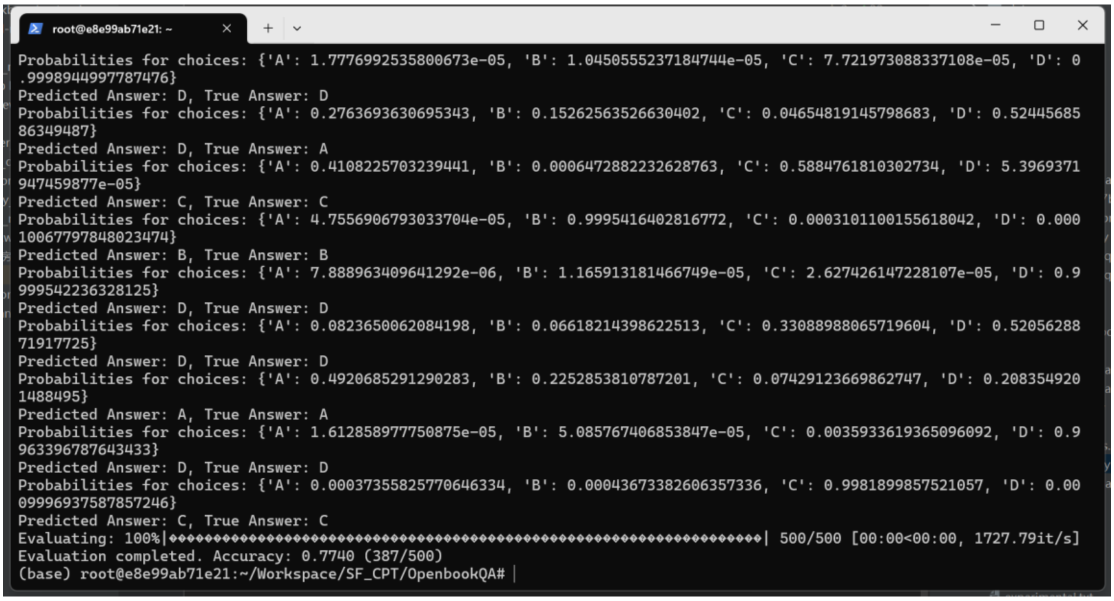
```
Predicted Answer: B, True Answer: A
Evaluating:  99%|██████████| 494/500 [04:16<00:03,  1.84it/s]
Probabilities for choices: {'A': 0.20556640625, 'B': 0.04693603515625, 'C': 0.7119140625, 'D': 0.03570
556640625}
Predicted Answer: C, True Answer: C
Evaluating:  99%|██████████| 495/500 [04:17<00:02,  1.89it/s]
Probabilities for choices: {'A': 0.040985107421875, 'B': 0.86279296875, 'C': 0.0625, 'D': 0.0334777832
03125}
Predicted Answer: B, True Answer: B
Evaluating:  99%|██████████| 496/500 [04:17<00:02,  1.93it/s]
Probabilities for choices: {'A': 0.0926513671875, 'B': 0.0238037109375, 'C': 0.0179595947265625, 'D':
0.86572265625}
Predicted Answer: D, True Answer: D
Evaluating:  99%|██████████| 497/500 [04:18<00:01,  1.97it/s]
Probabilities for choices: {'A': 0.128173828125, 'B': 0.2147216796875, 'C': 0.216430664625, 'D': 0.44
0673828125}
Predicted Answer: D, True Answer: D
Evaluating: 100%|██████████| 498/500 [04:18<00:01,  1.85it/s]
Probabilities for choices: {'A': 0.30322265625, 'B': 0.28271484375, 'C': 0.200439453125, 'D': 0.213378
90625}
Predicted Answer: A, True Answer: A
Evaluating: 100%|██████████| 499/500 [04:19<00:00,  1.91it/s]
Probabilities for choices: {'A': 0.01346588134765625, 'B': 0.00997161865234375, 'C': 0.366943359375,
'D': 0.60986328125}
Predicted Answer: D, True Answer: D
Evaluating: 100%|██████████| 500/500 [04:19<00:00,  1.92it/s]
Probabilities for choices: {'A': 0.048980712890625, 'B': 0.06646728515625, 'C': 0.8032265625, 'D': 0.
0814208984375}
Predicted Answer: C, True Answer: C
Evaluation completed. Accuracy: 0.8000 (400/500)
Logits saved!
```

Proxy tuning acc: **0.7080**



```
Probabilities for choices: {'A': 0.007318224757909775, 'B': 0.0021297249477356672, 'C': 0.009323660284280777, 'D': 0.981
2284111976624}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.20777848362922668, 'B': 0.16695469617843628, 'C': 0.1496572196483612, 'D': 0.47560966
01486206}
Predicted Answer: D, True Answer: A
Probabilities for choices: {'A': 0.05474036931991577, 'B': 0.008200244046747684, 'C': 0.9331247806549072, 'D': 0.0039345
20296752453}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 0.005237536504864693, 'B': 0.9673895239830017, 'C': 0.022750820964574814, 'D': 0.004622
109699994326}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 0.013443758711218834, 'B': 0.009096502326642889, 'C': 0.012629242613911629, 'D': 0.96483
05177688599}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.1276844292879l046, 'B': 0.22234925627708435, 'C': 0.31356269121170044, 'D': 0.3364036
6792678833}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.20074135065078735, 'B': 0.22746974229812622, 'C': 0.24595344066619873, 'D': 0.3258353
7697792053}
Predicted Answer: D, True Answer: A
Probabilities for choices: {'A': 0.003458289662376046, 'B': 0.0031735049560666084, 'C': 0.9366752505302429, 'D': 0.05669
290944933891}
Predicted Answer: C, True Answer: D
Probabilities for choices: {'A': 0.028618600219488144, 'B': 0.04645257443189621, 'C': 0.8495288491249084, 'D': 0.0753999
650478363}
Predicted Answer: C, True Answer: C
Evaluating: 100%|��������������������������������������������������������������������| 500/500 [00:00<00:00, 1722.12it/s]
Evaluation completed. Accuracy: 0.7080 (354/500)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/OpenbookQA#
```

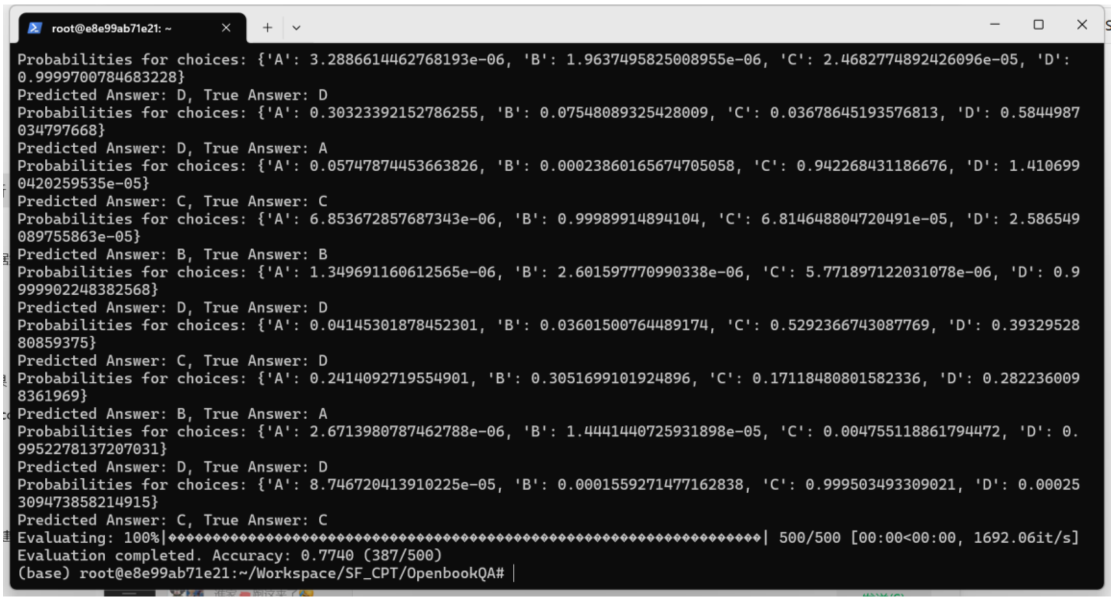
Consistent proxy tuning acc: **0.7740**



```
Probabilities for choices: {'A': 1.7776992535800673e-05, 'B': 1.0450555237184744e-05, 'C': 7.721973088337108e-05, 'D': 0
.9998944997787476}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.2763693630695343, 'B': 0.15262563526630402, 'C': 0.04654819145798683, 'D': 0.52445685
86349487}
Predicted Answer: D, True Answer: A
Probabilities for choices: {'A': 0.4108225703239441, 'B': 0.0006472882232628763, 'C': 0.5884761810302734, 'D': 5.3969371
947459877e-05}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 4.7556906793033704e-05, 'B': 0.9995416402816772, 'C': 0.0003101100155618042, 'D': 0.000
10067797848023474}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 7.888963409641292e-06, 'B': 1.165913181466749e-05, 'C': 2.627426147228107e-05, 'D': 0.9
999542236328125}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.0823650062084198, 'B': 0.06618214398622513, 'C': 0.33088988065719604, 'D': 0.52056288
71917725}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.4920685291290283, 'B': 0.2252853810787201, 'C': 0.07429123669862747, 'D': 0.208354920
1488495}
Predicted Answer: A, True Answer: A
Probabilities for choices: {'A': 1.612858977750875e-05, 'B': 5.085767406853847e-05, 'C': 0.0035933619365096092, 'D': 0.9
963396787643433}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.00037355825770646334, 'B': 0.00043673382606357336, 'C': 0.9981899857521057, 'D': 0.00
09996937587857246}
Predicted Answer: C, True Answer: C
Evaluating: 100%|��������������������������������������������������������������������| 500/500 [00:00<00:00, 1727.79it/s]
Evaluation completed. Accuracy: 0.7740 (387/500)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/OpenbookQA#
```

Gaussian process tuning acc: **0.7740**

400 / 4957 (8.07%)

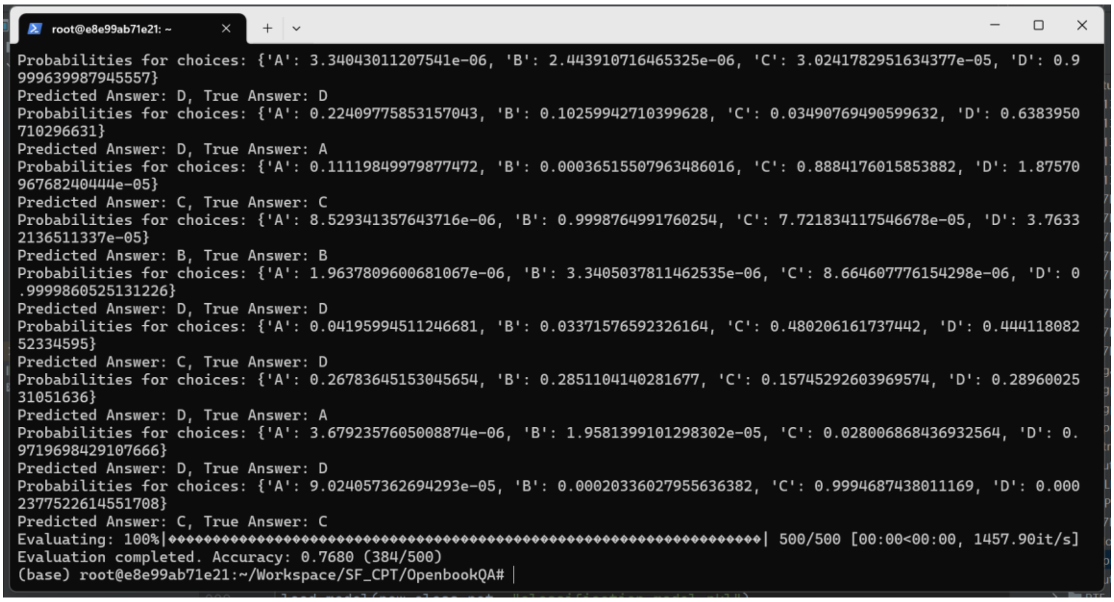


```
Probabilities for choices: {'A': 3.2886614462768193e-06, 'B': 1.9637495825008955e-06, 'C': 2.4682774892426096e-05, 'D':
0.9999700784683228}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.30323392152786255, 'B': 0.07548089325428009, 'C': 0.03678645193576813, 'D': 0.5844987
034797668}
Predicted Answer: D, True Answer: A
Probabilities for choices: {'A': 0.05747874453663826, 'B': 0.00023860165674705058, 'C': 0.942268431186676, 'D': 1.410699
0420259535e-05}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 6.853672857687343e-06, 'B': 0.99989914894104, 'C': 6.814648804720491e-05, 'D': 2.586549
089755863e-05}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 1.349691160612565e-06, 'B': 2.601597770990338e-06, 'C': 5.771897122031078e-06, 'D': 0.9
999902248382568}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.04145301878452301, 'B': 0.03601500764489174, 'C': 0.5292366743087769, 'D': 0.39329528
80859375}
Predicted Answer: C, True Answer: D
Probabilities for choices: {'A': 0.2414092719554901, 'B': 0.3051699101924896, 'C': 0.17118480801582336, 'D': 0.282236009
8361969}
Predicted Answer: B, True Answer: A
Probabilities for choices: {'A': 2.6713980787462788e-06, 'B': 1.4441440725931898e-05, 'C': 0.004755118861794472, 'D': 0.
9952278137207031}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 8.746720413910225e-05, 'B': 0.0001559271477162838, 'C': 0.999503493309021, 'D': 0.00025
309473858214915}
Predicted Answer: C, True Answer: C
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 500/500 [00:00<00:00, 1692.06it/s]
Evaluation completed. Accuracy: 0.7740 (387/500)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/OpenbookQA# |
```

Gaussian process with filter tuning acc: **0.7680**

Config: input_threshold=0.175, output_threshold=3

121 / 4957 (2.44%)



```
Probabilities for choices: {'A': 3.340430112075541e-06, 'B': 2.443910716465325e-06, 'C': 3.0241782951634377e-05, 'D': 0.9
999639987945557}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.22409775853157043, 'B': 0.10259942710399628, 'C': 0.03490769490599632, 'D': 0.6383950
710296631}
Predicted Answer: D, True Answer: A
Probabilities for choices: {'A': 0.11119849979877472, 'B': 0.00036515507963486016, 'C': 0.8884176015853882, 'D': 1.87570
96768240444e-05}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 8.529341357643716e-06, 'B': 0.9998764991760254, 'C': 7.721834117546678e-05, 'D': 3.7633
2136511337e-05}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 1.9637809600681067e-06, 'B': 3.3405037811462535e-06, 'C': 8.664607776154298e-06, 'D': 0
.9999860525131226}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 0.04195994511246681, 'B': 0.03371576592326164, 'C': 0.480206161737442, 'D': 0.444118082
52334595}
Predicted Answer: C, True Answer: D
Probabilities for choices: {'A': 0.26783645153045654, 'B': 0.2851104140281677, 'C': 0.15745292603969574, 'D': 0.28960025
31051636}
Predicted Answer: D, True Answer: A
Probabilities for choices: {'A': 3.6792357605008874e-06, 'B': 1.9581399101298302e-05, 'C': 0.028006868436932564, 'D': 0.
9719698429107666}
Predicted Answer: D, True Answer: D
Probabilities for choices: {'A': 9.024057362694293e-05, 'B': 0.00020336027955636382, 'C': 0.9994687438011169, 'D': 0.000
2377522614551708}
Predicted Answer: C, True Answer: C
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 500/500 [00:00<00:00, 1457.90it/s]
Evaluation completed. Accuracy: 0.7680 (384/500)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/OpenbookQA# |
```