benchmark

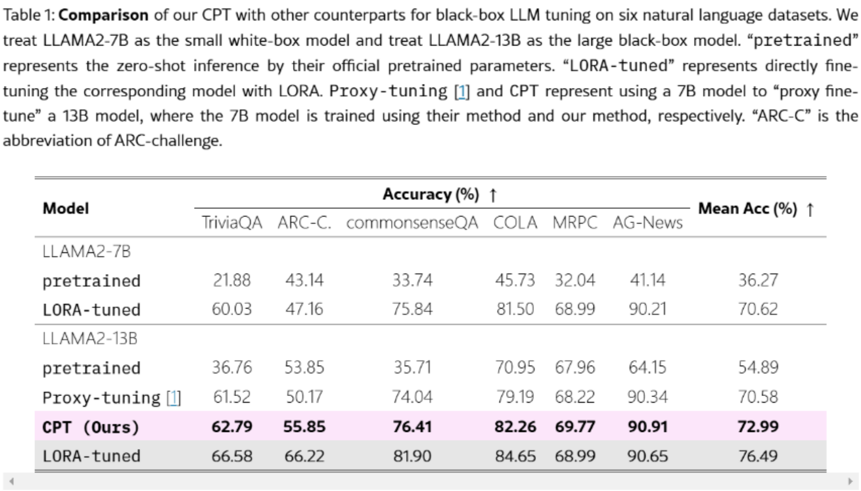
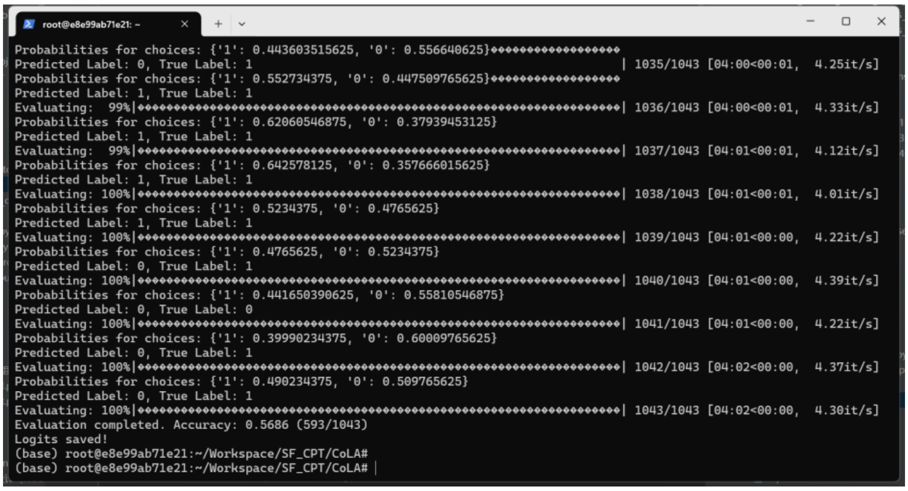


Table 1: **Comparison** of our CPT with other counterparts for black-box LLM tuning on six natural language datasets. We treat LLAMA2-7B as the small white-box model and treat LLAMA2-13B as the large black-box model. "`pretrained`" represents the zero-shot inference by their official pretrained parameters. "`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.

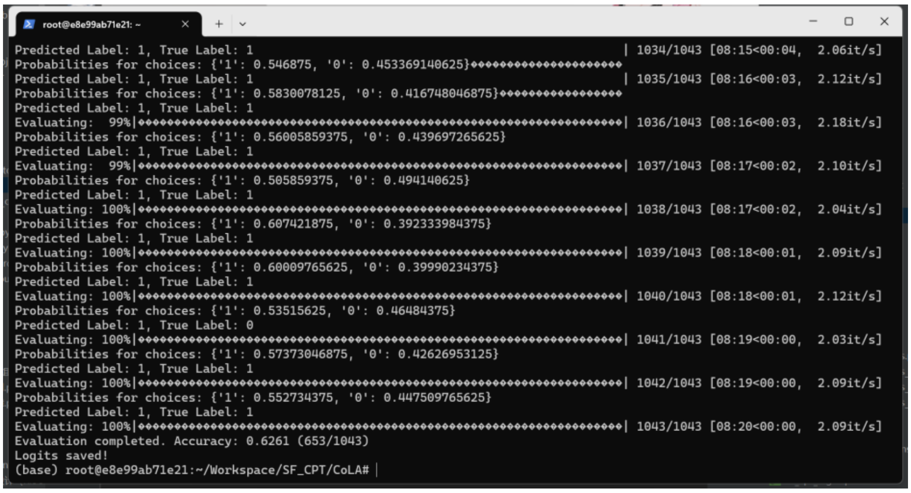
Model	**Accuracy (%) ↑**						**Mean Acc (%) ↑**
	TriviaQA	ARC-C.	commonsenseQA	COLA	MRPC	AG-News	
LLAMA2-7B							
`pretrained`	21.88	43.14	33.74	45.73	32.04	41.14	36.27
`LORA-tuned`	60.03	47.16	75.84	81.50	68.99	90.21	70.62
LLAMA2-13B							
`pretrained`	36.76	53.85	35.71	70.95	67.96	64.15	54.89
`Proxy-tuning` [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 acc: **0.5686**



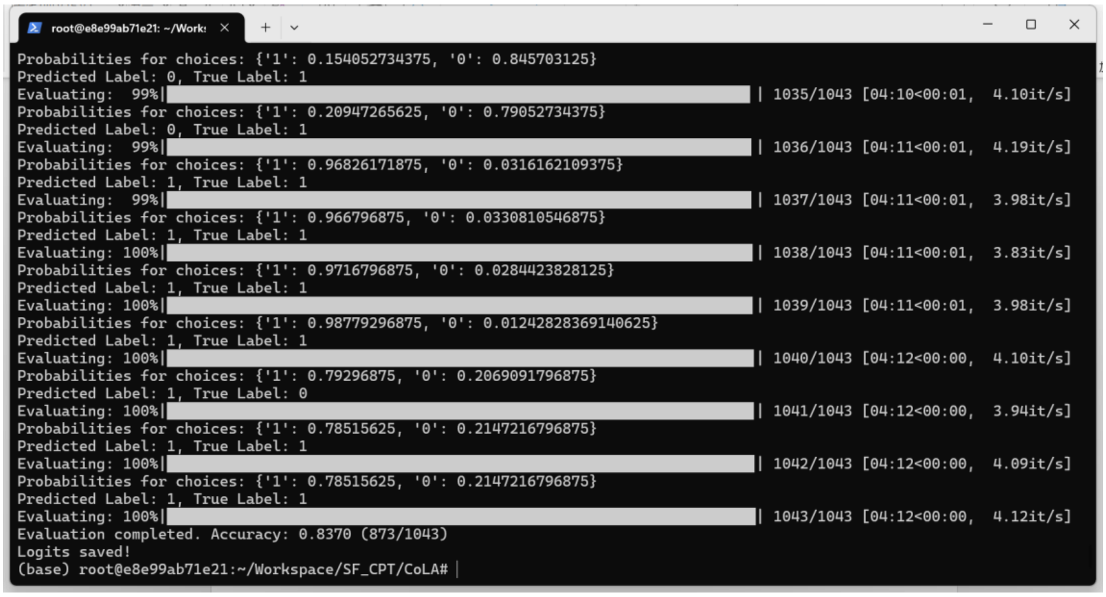
```
Probabilities for choices: {'1': 0.443603515625, '0': 0.556640625}◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆
Predicted Label: 0, True Label: 1                                                   | 1035/1043 [04:00<00:01,  4.25it/s]
Probabilities for choices: {'1': 0.552734375, '0': 0.447509765625}◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆
Predicted Label: 1, True Label: 1
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1036/1043 [04:00<00:01,  4.33it/s]
Probabilities for choices: {'1': 0.62060546875, '0': 0.37939453125}
Predicted Label: 1, True Label: 1
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1037/1043 [04:01<00:01,  4.12it/s]
Probabilities for choices: {'1': 0.642578125, '0': 0.357666015625}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1038/1043 [04:01<00:01,  4.01it/s]
Probabilities for choices: {'1': 0.5234375, '0': 0.4765625}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1039/1043 [04:01<00:00,  4.22it/s]
Probabilities for choices: {'1': 0.4765625, '0': 0.5234375}
Predicted Label: 0, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1040/1043 [04:01<00:00,  4.39it/s]
Probabilities for choices: {'1': 0.441650390625, '0': 0.55810546875}
Predicted Label: 0, True Label: 0
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1041/1043 [04:01<00:00,  4.22it/s]
Probabilities for choices: {'1': 0.39990234375, '0': 0.60009765625}
Predicted Label: 0, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1042/1043 [04:02<00:00,  4.37it/s]
Probabilities for choices: {'1': 0.490234375, '0': 0.509765625}
Predicted Label: 0, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1043/1043 [04:02<00:00,  4.30it/s]
Evaluation completed. Accuracy: 0.5686 (593/1043)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CoLA#
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CoLA# |
```

13b pretrain acc: **0.6261**



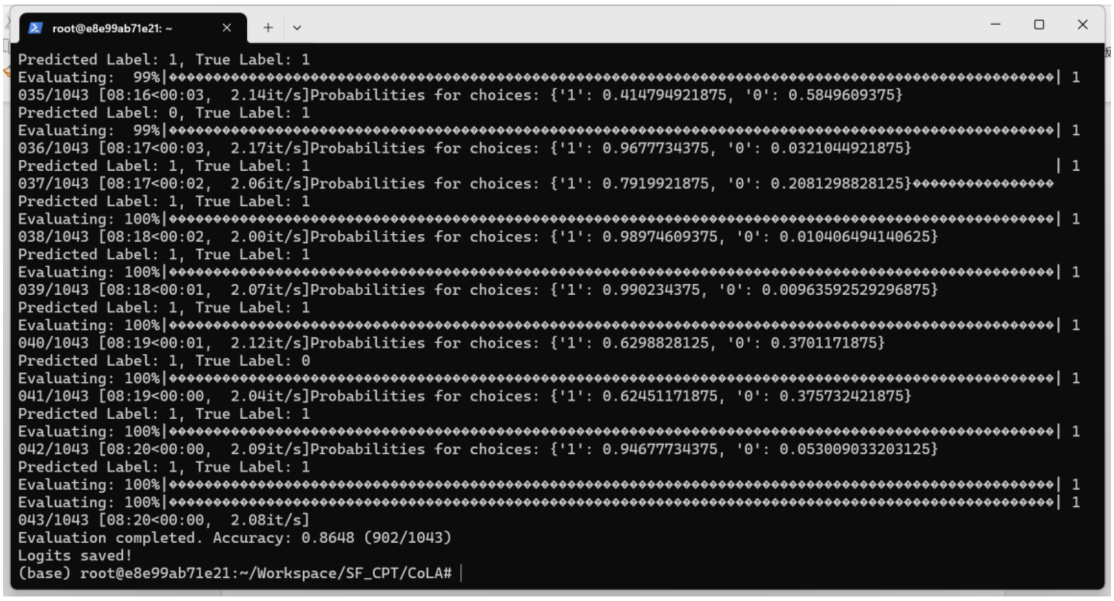
```
Predicted Label: 1, True Label: 1                                                   | 1034/1043 [08:15<00:04,  2.06it/s]
Probabilities for choices: {'1': 0.546875, '0': 0.453369140625}◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆
Predicted Label: 1, True Label: 1                                                   | 1035/1043 [08:16<00:03,  2.12it/s]
Probabilities for choices: {'1': 0.5830078125, '0': 0.416748046875}◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆
Predicted Label: 1, True Label: 1
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1036/1043 [08:16<00:03,  2.18it/s]
Probabilities for choices: {'1': 0.56005859375, '0': 0.439697265625}
Predicted Label: 1, True Label: 1
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1037/1043 [08:17<00:02,  2.10it/s]
Probabilities for choices: {'1': 0.505859375, '0': 0.494140625}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1038/1043 [08:17<00:02,  2.04it/s]
Probabilities for choices: {'1': 0.607421875, '0': 0.392333984375}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1039/1043 [08:18<00:01,  2.09it/s]
Probabilities for choices: {'1': 0.60009765625, '0': 0.39990234375}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1040/1043 [08:18<00:01,  2.12it/s]
Probabilities for choices: {'1': 0.53515625, '0': 0.46484375}
Predicted Label: 1, True Label: 0
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1041/1043 [08:19<00:00,  2.03it/s]
Probabilities for choices: {'1': 0.57373046875, '0': 0.42626953125}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1042/1043 [08:19<00:00,  2.09it/s]
Probabilities for choices: {'1': 0.552734375, '0': 0.447509765625}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1043/1043 [08:20<00:00,  2.09it/s]
Evaluation completed. Accuracy: 0.6261 (653/1043)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CoLA# |
```

7b lora-tune acc: **0.8370**



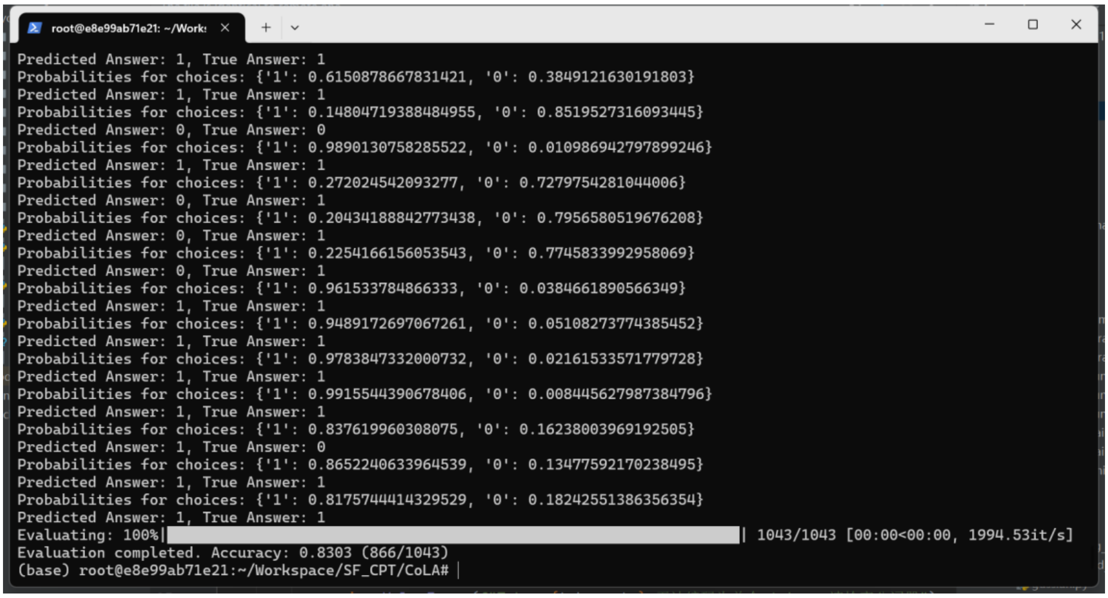
```
Probabilities for choices: {'1': 0.154052734375, '0': 0.845703125}
Predicted Label: 0, True Label: 1
Evaluating:  99%|█████████████████████████████████████████████████████████████████| 1035/1043 [04:10<00:01,  4.10it/s]
Probabilities for choices: {'1': 0.20947265625, '0': 0.79052734375}
Predicted Label: 0, True Label: 1
Evaluating:  99%|█████████████████████████████████████████████████████████████████| 1036/1043 [04:11<00:01,  4.19it/s]
Probabilities for choices: {'1': 0.96826171875, '0': 0.0316162109375}
Predicted Label: 1, True Label: 1
Evaluating:  99%|█████████████████████████████████████████████████████████████████| 1037/1043 [04:11<00:01,  3.98it/s]
Probabilities for choices: {'1': 0.966796875, '0': 0.0330810546875}
Predicted Label: 1, True Label: 1
Evaluating: 100%|█████████████████████████████████████████████████████████████████| 1038/1043 [04:11<00:01,  3.83it/s]
Probabilities for choices: {'1': 0.9716796875, '0': 0.0284423828125}
Predicted Label: 1, True Label: 1
Evaluating: 100%|█████████████████████████████████████████████████████████████████| 1039/1043 [04:11<00:01,  3.98it/s]
Probabilities for choices: {'1': 0.98779296875, '0': 0.01242828369140625}
Predicted Label: 1, True Label: 1
Evaluating: 100%|█████████████████████████████████████████████████████████████████| 1040/1043 [04:12<00:00,  4.10it/s]
Probabilities for choices: {'1': 0.79296875, '0': 0.2069091796875}
Predicted Label: 1, True Label: 0
Evaluating: 100%|█████████████████████████████████████████████████████████████████| 1041/1043 [04:12<00:00,  3.94it/s]
Probabilities for choices: {'1': 0.78515625, '0': 0.2147216796875}
Predicted Label: 1, True Label: 1
Evaluating: 100%|█████████████████████████████████████████████████████████████████| 1042/1043 [04:12<00:00,  4.09it/s]
Probabilities for choices: {'1': 0.78515625, '0': 0.2147216796875}
Predicted Label: 1, True Label: 1
Evaluating: 100%|█████████████████████████████████████████████████████████████████| 1043/1043 [04:12<00:00,  4.12it/s]
Evaluation completed. Accuracy: 0.8370 (873/1043)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CoLA# 
```

13b lora-tune acc: **0.8648**



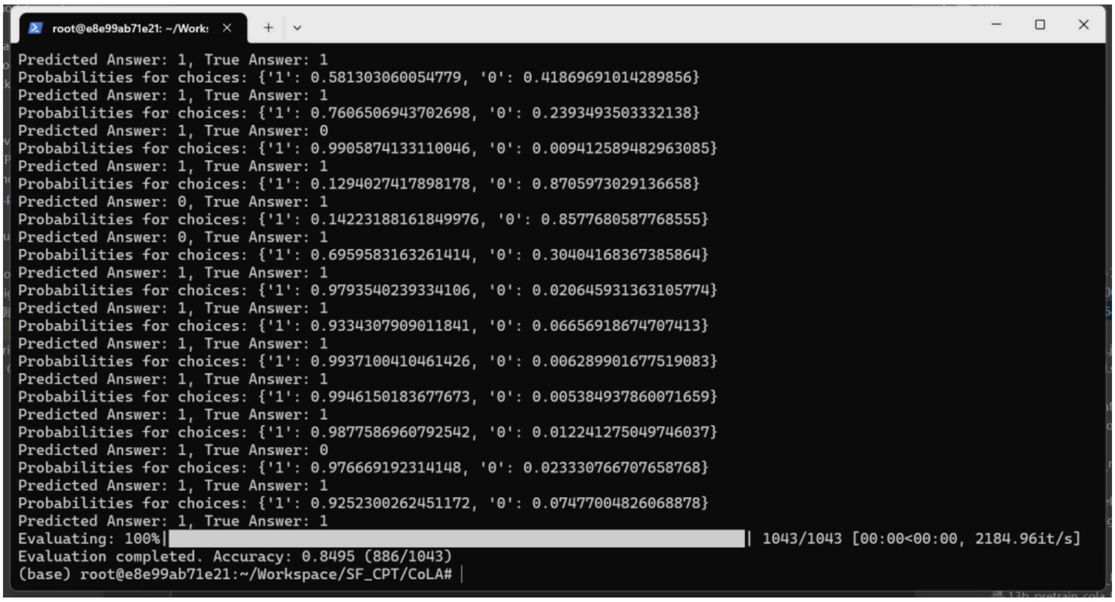
```
Predicted Label: 1, True Label: 1
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1
035/1043 [08:16<00:03,  2.14it/s]Probabilities for choices: {'1': 0.414794921875, '0': 0.5849609375}
Predicted Label: 0, True Label: 1
Evaluating:  99%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1
036/1043 [08:17<00:03,  2.17it/s]Probabilities for choices: {'1': 0.9677734375, '0': 0.0321044921875}
Predicted Label: 1, True Label: 1                                                                     | 1
037/1043 [08:17<00:02,  2.06it/s]Probabilities for choices: {'1': 0.7919921875, '0': 0.2081298828125}◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1
038/1043 [08:18<00:02,  2.00it/s]Probabilities for choices: {'1': 0.98974609375, '0': 0.010406494140625}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1
039/1043 [08:18<00:01,  2.07it/s]Probabilities for choices: {'1': 0.990234375, '0': 0.00963592529296875}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1
040/1043 [08:19<00:01,  2.12it/s]Probabilities for choices: {'1': 0.6298828125, '0': 0.3701171875}
Predicted Label: 1, True Label: 0
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1
041/1043 [08:19<00:00,  2.04it/s]Probabilities for choices: {'1': 0.62451171875, '0': 0.375732421875}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1
042/1043 [08:20<00:00,  2.09it/s]Probabilities for choices: {'1': 0.94677734375, '0': 0.053009033203125}
Predicted Label: 1, True Label: 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1
Evaluating: 100%|◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆| 1
043/1043 [08:20<00:00,  2.08it/s]
Evaluation completed. Accuracy: 0.8648 (902/1043)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CoLA# 
```

Proxy tuning: **0.8303**



```
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.6150878667831421, '0': 0.3849121630191803}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.14804719388484955, '0': 0.8519527316093445}
Predicted Answer: 0, True Answer: 0
Probabilities for choices: {'1': 0.9890130758285522, '0': 0.010986942797899246}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.272024542093277, '0': 0.727975428104406}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.20434188842773438, '0': 0.7956580519676208}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.2254166156053543, '0': 0.7745833992958069}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.961533784866333, '0': 0.0384661890566349}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9489172697067261, '0': 0.05108273774385452}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9783847332000732, '0': 0.02161533571779728}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9915544390678406, '0': 0.008445627987384796}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.837619960308075, '0': 0.16238003969192505}
Predicted Answer: 1, True Answer: 0
Probabilities for choices: {'1': 0.8652240633964539, '0': 0.13477592170238495}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.8175744414329529, '0': 0.18242551386356354}
Predicted Answer: 1, True Answer: 1
Evaluating: 100%|██████████████████████████████████████████████████| 1043/1043 [00:00<00:00, 1994.53it/s]
Evaluation completed. Accuracy: 0.8303 (866/1043)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CoLA#
```

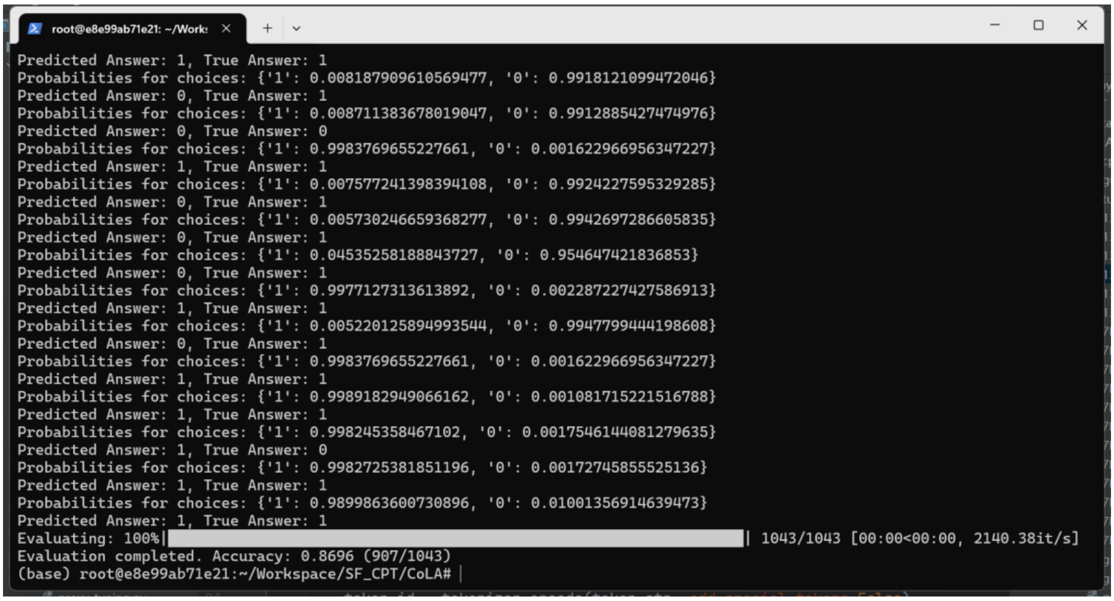
Consistent proxy tuning: **0.8495**



```
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.581303060054779, '0': 0.41869691014289856}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.7606506943702698, '0': 0.2393493503332138}
Predicted Answer: 1, True Answer: 0
Probabilities for choices: {'1': 0.9905874133110046, '0': 0.009412589482963085}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.1294027417898178, '0': 0.8705973029136658}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.14223188161849976, '0': 0.8577680587768555}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.6959583163261414, '0': 0.30404168367385864}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9793540239334106, '0': 0.020645931363105774}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9334307909011841, '0': 0.06656918674707413}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9937100410461426, '0': 0.006289901677519083}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9946150183677673, '0': 0.005384937860071659}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9877586960792542, '0': 0.012241275049746037}
Predicted Answer: 1, True Answer: 0
Probabilities for choices: {'1': 0.976669192314148, '0': 0.02333076670765876 8}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9252300262451172, '0': 0.07477004826068878}
Predicted Answer: 1, True Answer: 1
Evaluating: 100%|██████████████████████████████████████████████████| 1043/1043 [00:00<00:00, 2184.96it/s]
Evaluation completed. Accuracy: 0.8495 (886/1043)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CoLA#
```

GP tune acc: **0.8696**

500 / 8551 (5.85%)

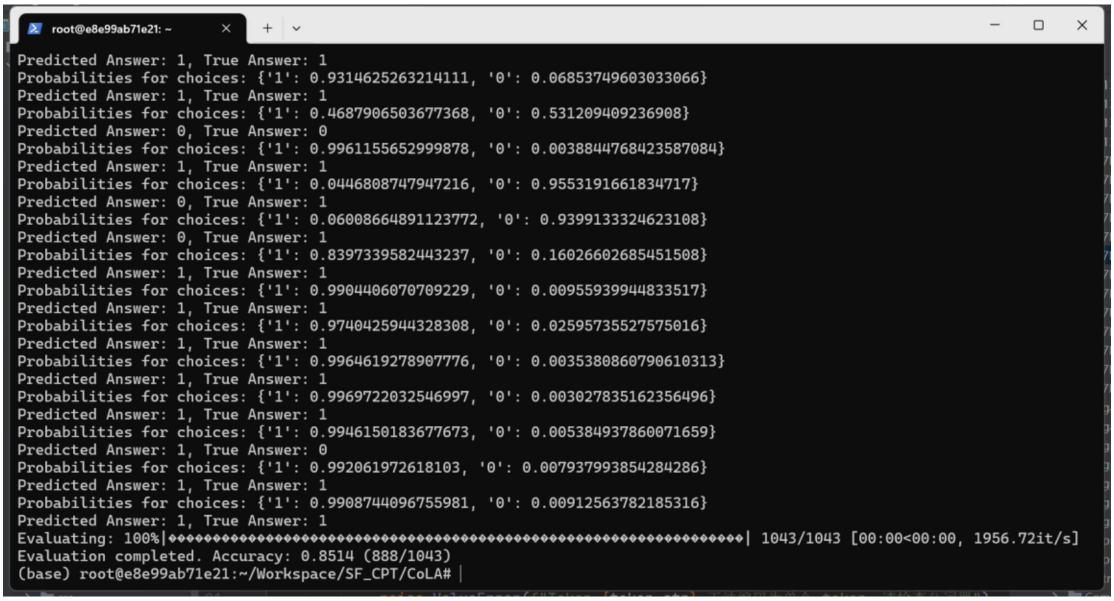


```
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.008187909610569477, '0': 0.9918121099472046}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.008711383678019047, '0': 0.9912885427474976}
Predicted Answer: 0, True Answer: 0
Probabilities for choices: {'1': 0.9983769655227661, '0': 0.001622966956347227}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.007577241398394108, '0': 0.9924227595329285}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.005730246659368277, '0': 0.9942697286605835}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.04535258188843727, '0': 0.954647421836853}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.9977127313613892, '0': 0.002287227427586913}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.005220125894993544, '0': 0.9947799444198608}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.9983769655227661, '0': 0.001622966956347227}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9989182949066162, '0': 0.001081715221516788}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.998245358467102, '0': 0.0017546144081279635}
Predicted Answer: 1, True Answer: 0
Probabilities for choices: {'1': 0.9982725381851196, '0': 0.00172745855525136}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9899863600730896, '0': 0.01001356914639473}
Predicted Answer: 1, True Answer: 1
Evaluating: 100%|████████████████████████████████████████████████████████████████| 1043/1043 [00:00<00:00, 2140.38it/s]
Evaluation completed. Accuracy: 0.8696 (907/1043)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CoLA#
```

GP with filter accuracy: **0.8514**

Conifg: input_threshold=0.075, output_threshold=1.5

171/8551　（1.78%）



```
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9314625263214111, '0': 0.06853749603033066}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.4687906503677368, '0': 0.531209409236908}
Predicted Answer: 0, True Answer: 0
Probabilities for choices: {'1': 0.9961155652999878, '0': 0.0038844768423587084}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.0446808747947216, '0': 0.9553191661834717}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.06008664891123772, '0': 0.9399133324623108}
Predicted Answer: 0, True Answer: 1
Probabilities for choices: {'1': 0.8397339582443237, '0': 0.16026602685451508}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9904406070709229, '0': 0.00955939944833517}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9740425944328308, '0': 0.02595735527575016}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9964619278907776, '0': 0.0035380860790610313}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9969722032546997, '0': 0.003027835162356496}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9946150183677673, '0': 0.005384937860071659}
Predicted Answer: 1, True Answer: 0
Probabilities for choices: {'1': 0.992061972618103, '0': 0.007937993854284286}
Predicted Answer: 1, True Answer: 1
Probabilities for choices: {'1': 0.9908744096755981, '0': 0.00912563782185316}
Predicted Answer: 1, True Answer: 1
Evaluating: 100%|����������������������������������������������������������������| 1043/1043 [00:00<00:00, 1956.72it/s]
Evaluation completed. Accuracy: 0.8514 (888/1043)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/CoLA#
```