

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 (%) $\uparrow$						Mean Acc (%) $\uparrow$
	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
<b>CPT (Ours)</b>	<b>62.79</b>	<b>55.85</b>	<b>76.41</b>	<b>82.26</b>	<b>69.77</b>	<b>90.91</b>	<b>72.99</b>
LORA-tuned	66.58	66.22	81.90	84.65	68.99	90.65	76.49

benchmark

7b pretrain accuracy: **0.3712**

```

root@e8e99ab71e21: ~/Work:
13637542724609375}
Predicted Answer: A, True Answer: D
Evaluating: 98% | 293/299 [01:16<00:01, 3.99it/s]
Probabilities for choices: {'A': 0.27685546875, 'B': 0.18017578125, 'C': 0.347412109375, 'D': 0.193359375, 'E': 0.002286
9110107421875}
Predicted Answer: C, True Answer: D
Evaluating: 98% | 294/299 [01:16<00:01, 3.93it/s]
Probabilities for choices: {'A': 0.22412109375, 'B': 0.260009765625, 'C': 0.316162109375, 'D': 0.19775390625, 'E': 0.002
0313262939453125}
Predicted Answer: C, True Answer: C
Evaluating: 99% | 295/299 [01:16<00:00, 4.16it/s]
Probabilities for choices: {'A': 0.2479248046875, 'B': 0.384033203125, 'C': 0.2039794921875, 'D': 0.16259765625, 'E': 0.
0013427734375}
Predicted Answer: B, True Answer: B
Evaluating: 99% | 296/299 [01:16<00:00, 4.02it/s]
Probabilities for choices: {'A': 0.251220703125, 'B': 0.1673583984375, 'C': 0.29833984375, 'D': 0.2802734375, 'E': 0.002
7904510498046875}
Predicted Answer: C, True Answer: C
Evaluating: 99% | 297/299 [01:17<00:00, 3.93it/s]
Probabilities for choices: {'A': 0.1787109375, 'B': 0.328857421875, 'C': 0.390380859375, 'D': 0.10186767578125, 'E': 0.0
002894401550292969}
Predicted Answer: C, True Answer: D
Evaluating: 100% | 298/299 [01:17<00:00, 3.88it/s]
Probabilities for choices: {'A': 0.3251953125, 'B': 0.34619140625, 'C': 0.12939453125, 'D': 0.19873046875, 'E': 0.000627
5177001953125}
Predicted Answer: B, True Answer: D
Evaluating: 100% | 299/299 [01:17<00:00, 3.85it/s]
Evaluation completed. Accuracy: 0.3712 (111/299)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/ARC-C#

```

13b pretrain accuracy: **0.5351**

```

root@e8e99ab71e21: ~/Work:
084075927734375}
Predicted Answer: A, True Answer: D
Evaluating: 98% | 293/299 [02:34<00:03, 1.91it/s]
Probabilities for choices: {'A': 0.236572265625, 'B': 0.289794921875, 'C': 0.31591796875, 'D': 0.1446533203125, 'E': 0.0
129852294921875}
Predicted Answer: C, True Answer: D
Evaluating: 98% | 294/299 [02:35<00:02, 1.90it/s]
Probabilities for choices: {'A': 0.2208251953125, 'B': 0.1219482421875, 'C': 0.33154296875, 'D': 0.311279296875, 'E': 0.
01450347900390625}
Predicted Answer: C, True Answer: C
Evaluating: 99% | 295/299 [02:35<00:01, 2.01it/s]
Probabilities for choices: {'A': 0.147705078125, 'B': 0.68798828125, 'C': 0.102294921875, 'D': 0.058746337890625, 'E': 0.
0030765533447265625}
Predicted Answer: B, True Answer: B
Evaluating: 99% | 296/299 [02:36<00:01, 1.97it/s]
Probabilities for choices: {'A': 0.314697265625, 'B': 0.145263671875, 'C': 0.2470703125, 'D': 0.2734375, 'E': 0.01965332
03125}
Predicted Answer: A, True Answer: C
Evaluating: 99% | 297/299 [02:36<00:01, 1.94it/s]
Probabilities for choices: {'A': 0.036407470703125, 'B': 0.5224609375, 'C': 0.1212158203125, 'D': 0.314453125, 'E': 0.00
5558013916015625}
Predicted Answer: B, True Answer: D
Evaluating: 100% | 298/299 [02:37<00:00, 1.92it/s]
Probabilities for choices: {'A': 0.260009765625, 'B': 0.297119140625, 'C': 0.2259521484375, 'D': 0.208984375, 'E': 0.007
762908935546875}
Predicted Answer: B, True Answer: D
Evaluating: 100% | 299/299 [02:38<00:00, 1.89it/s]
Evaluation completed. Accuracy: 0.5351 (160/299)
Logits saved!
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/ARC-C#

```

7b lora-tune accuracy: **0.4649**

```
root@e8e99ab71e21: ~  
Predicted Answer: D, True Answer: D  
Evaluating: 98%|██████████| 293/299 [01:15<00:01, 3.99it/s] Probabilities for choices: {'A': 0.177001953125, 'B': 0.19140625, 'C': 0.383544921875, 'D': 0.24755859375, 'E': 0.0005502700805664062}  
Predicted Answer: C, True Answer: D  
Evaluating: 98%|██████████| 294/299 [01:16<00:01, 3.95it/s] Probabilities for choices: {'A': 0.186279296875, 'B': 0.146240234375, 'C': 0.3046875, 'D': 0.36181640625, 'E': 0.0008625984191894531}  
Predicted Answer: D, True Answer: C  
Evaluating: 99%|██████████| 295/299 [01:16<00:00, 4.19it/s] Probabilities for choices: {'A': 0.17041015625, 'B': 0.68994140625, 'C': 0.109130859375, 'D': 0.030548095703125, 'E': 0.000148773193359375}  
Predicted Answer: B, True Answer: B  
Evaluating: 99%|██████████| 296/299 [01:16<00:00, 4.07it/s] Probabilities for choices: {'A': 0.1197509765625, 'B': 0.152587890625, 'C': 0.4736328125, 'D': 0.253662109375, 'E': 0.0004253387451171875}  
Predicted Answer: C, True Answer: C  
Evaluating: 99%|██████████| 297/299 [01:16<00:00, 4.01it/s] Probabilities for choices: {'A': 0.1240234375, 'B': 0.464599609375, 'C': 0.2467041015625, 'D': 0.164306640625, 'E': 0.0003933906555175781}  
Predicted Answer: B, True Answer: D  
Evaluating: 100%|██████████| 298/299 [01:17<00:00, 3.96it/s] Probabilities for choices: {'A': 0.267578125, 'B': 0.3330078125, 'C': 0.24169921875, 'D': 0.1573486328125, 'E': 0.0004315376281738281}  
Predicted Answer: B, True Answer: D  
Evaluating: 100%|██████████| 299/299 [01:17<00:00, 3.86it/s]  
Evaluation completed. Accuracy: 0.4649 (139/299)  
Logits saved!  
(base) root@e8e99ab71e21: ~/Workspace/SF_CPT/ARC-C#
```

13b lora-tune accuracy: **0.5619**

```
Predicted Answer: C, True Answer: C  
Evaluating: 98%|██████████| 293/299 [03:01<00:03, 1.69it/s]  
Probabilities for choices: {'A': 0.378173828125, 'B': 0.202392578125, 'C': 0.1947021484375, 'D': 0.2205810546875, 'E': 0.004039764404296875}  
Predicted Answer: A, True Answer: D  
Evaluating: 98%|██████████| 294/299 [03:02<00:02, 1.67it/s]  
Probabilities for choices: {'A': 0.2301025390625, 'B': 0.3046875, 'C': 0.29541015625, 'D': 0.1669921875, 'E': 0.0028285980224609375}  
Predicted Answer: B, True Answer: D  
Evaluating: 99%|██████████| 295/299 [03:03<00:02, 1.77it/s]  
Probabilities for choices: {'A': 0.384033203125, 'B': 0.1317138671875, 'C': 0.182861328125, 'D': 0.28759765625, 'E': 0.0139312744140625}  
Predicted Answer: A, True Answer: C  
Evaluating: 99%|██████████| 296/299 [03:03<00:01, 1.72it/s]  
Probabilities for choices: {'A': 0.11859130859375, 'B': 0.75537109375, 'C': 0.04534912109375, 'D': 0.0789794921875, 'E': 0.001758575439453125}  
Predicted Answer: B, True Answer: B  
Evaluating: 99%|██████████| 297/299 [03:04<00:01, 1.69it/s]  
Probabilities for choices: {'A': 0.22607421875, 'B': 0.0745849609375, 'C': 0.428955078125, 'D': 0.268546875, 'E': 0.001753807067810938}  
Predicted Answer: C, True Answer: C  
Evaluating: 100%|██████████| 298/299 [03:04<00:00, 1.68it/s]  
Probabilities for choices: {'A': 0.040496826171875, 'B': 0.36962890625, 'C': 0.039886474609375, 'D': 0.54638671875, 'E': 0.003917694091796875}  
Predicted Answer: D, True Answer: D  
Evaluating: 100%|██████████| 299/299 [03:05<00:00, 1.61it/s]  
Probabilities for choices: {'A': 0.303955078125, 'B': 0.321044921875, 'C': 0.1380615234375, 'D': 0.23486328125, 'E': 0.002292633056640625}  
Predicted Answer: B, True Answer: D  
Evaluation completed. Accuracy: 0.5619 (168/299)  
Logits saved!
```

Proxy tuning accuracy: **0.5318**

```
root@e8e99ab71e21: ~  
Probabilities for choices: {'A': 0.40921252965927124, 'B': 0.21903569996356964, 'C': 0.22776107490062714, 'D': 0.14252923429012299, 'E': 0.0014615083346143365}  
Predicted Answer: A, True Answer: D  
Probabilities for choices: {'A': 0.027094759047031403, 'B': 0.05647032707929611, 'C': 0.749684751033783, 'D': 0.16597551107406616, 'E': 0.0007746474584564567}  
Predicted Answer: C, True Answer: C  
Probabilities for choices: {'A': 0.20120267570018768, 'B': 0.1314375102519989, 'C': 0.33432891964912415, 'D': 0.33172714710235596, 'E': 0.00130375730805099}  
Predicted Answer: C, True Answer: D  
Probabilities for choices: {'A': 0.15817074477672577, 'B': 0.2819698750972748, 'C': 0.3592389225959778, 'D': 0.19839058816432953, 'E': 0.00222989940084517}  
Predicted Answer: C, True Answer: D  
Probabilities for choices: {'A': 0.16637562215328217, 'B': 0.07211849838495255, 'C': 0.28747129440307617, 'D': 0.47027164697647095, 'E': 0.0037628833670169115}  
Predicted Answer: D, True Answer: C  
Probabilities for choices: {'A': 0.08715905249118805, 'B': 0.8531920313835144, 'C': 0.04889179766178131, 'D': 0.010532370768487453, 'E': 0.00022465195797849447}  
Predicted Answer: B, True Answer: B  
Probabilities for choices: {'A': 0.15247280895709991, 'B': 0.14549055695533752, 'C': 0.4343545734882355, 'D': 0.26551562547683716, 'E': 0.0021664283704012632}  
Predicted Answer: C, True Answer: C  
Probabilities for choices: {'A': 0.028591202571988106, 'B': 0.5544312596321106, 'C': 0.07987339794635773, 'D': 0.33366259932518005, 'E': 0.0034415144473314285}  
Predicted Answer: B, True Answer: D  
Probabilities for choices: {'A': 0.21059980988502502, 'B': 0.27682822942733765, 'C': 0.3554544746875763, 'D': 0.15407811105251312, 'E': 0.0030394557397812605}  
Predicted Answer: C, True Answer: D  
Evaluating: 100%|*****| 299/299 [00:00<00:00, 1564.57it/s]  
Evaluation completed. Accuracy: 0.5318 (159/299)  
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/ARC-C#
```

Consistent Proxy tuning accuracy: **0.5552**

```
root@e8e99ab71e21: ~/Work  
Probabilities for choices: {'A': 0.4515986740589142, 'B': 0.2513525187969208, 'C': 0.2149931937456131, 'D': 0.07034550607204437, 'E': 0.011710081249475479}  
Predicted Answer: A, True Answer: D  
Probabilities for choices: {'A': 0.08249140530824661, 'B': 0.07749350368976593, 'C': 0.6906920671463013, 'D': 0.14477692544460297, 'E': 0.004546040203422308}  
Predicted Answer: C, True Answer: C  
Probabilities for choices: {'A': 0.3895336389541626, 'B': 0.21512091159820557, 'C': 0.24186693131923676, 'D': 0.1372743546962738, 'E': 0.016204075887799263}  
Predicted Answer: A, True Answer: D  
Probabilities for choices: {'A': 0.21912842988967896, 'B': 0.3447374105453491, 'C': 0.2902979850769043, 'D': 0.12291990965604782, 'E': 0.02291623316705227}  
Predicted Answer: B, True Answer: D  
Probabilities for choices: {'A': 0.20518703758716583, 'B': 0.08756308257579803, 'C': 0.32406795024871826, 'D': 0.3545314073562622, 'E': 0.02865052968263626}  
Predicted Answer: D, True Answer: C  
Probabilities for choices: {'A': 0.10783005505800247, 'B': 0.7844096422195435, 'C': 0.06962031126022339, 'D': 0.034464556723833084, 'E': 0.0036753567401319742}  
Predicted Answer: B, True Answer: B  
Probabilities for choices: {'A': 0.33307650685310364, 'B': 0.1356305629014969, 'C': 0.22802743315696716, 'D': 0.2634851932525635, 'E': 0.03978031501173973}  
Predicted Answer: A, True Answer: C  
Probabilities for choices: {'A': 0.015290730632841587, 'B': 0.5224330425262451, 'C': 0.06000538542866707, 'D': 0.3882392644882202, 'E': 0.014031562954187393}  
Predicted Answer: B, True Answer: D  
Probabilities for choices: {'A': 0.22919870913028717, 'B': 0.27006208896636963, 'C': 0.2721801996231079, 'D': 0.20868779718875885, 'E': 0.019871307536959648}  
Predicted Answer: C, True Answer: D  
Evaluating: 100%|*****| 299/299 [00:00<00:00, 1568.42it/s]  
Evaluation completed. Accuracy: 0.5552 (166/299)  
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/ARC-C#
```

Gaussian Process accuracy: **0.5619**

150 / 1119 (13.40%)

```
root@e8e99ab71e21: ~/Work
Probabilities for choices: {'A': 0.4266722798347473, 'B': 0.2547776401042938, 'C': 0.2213546633720398, 'D': 0.0863457843
6613083, 'E': 0.010849741287529469}
Predicted Answer: A, True Answer: D
Probabilities for choices: {'A': 0.08680474758148193, 'B': 0.08612923324108124, 'C': 0.6617655754089355, 'D': 0.16091059
148311615, 'E': 0.004389811772853136}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 0.3630264103412628, 'B': 0.2184731364250183, 'C': 0.24182771146297455, 'D': 0.161091998
21949005, 'E': 0.015580779872834682}
Predicted Answer: A, True Answer: D
Probabilities for choices: {'A': 0.21166743338108063, 'B': 0.3382435739040375, 'C': 0.2961759567260742, 'D': 0.132458075
88100433, 'E': 0.021454915404319763}
Predicted Answer: B, True Answer: D
Probabilities for choices: {'A': 0.20601683855056763, 'B': 0.09432138502597809, 'C': 0.3116935193538666, 'D': 0.36157080
5311203, 'E': 0.026397529989480972}
Predicted Answer: D, True Answer: C
Probabilities for choices: {'A': 0.10745109617710114, 'B': 0.7755700945854187, 'C': 0.07442919909954071, 'D': 0.03906852
379441261, 'E': 0.003481100080534816}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 0.3170436918735504, 'B': 0.13850612938404083, 'C': 0.23195412755012512, 'D': 0.27761271
595954895, 'E': 0.03488330543041229}
Predicted Answer: A, True Answer: C
Probabilities for choices: {'A': 0.017200395464897156, 'B': 0.5026692748069763, 'C': 0.0712936669588089, 'D': 0.39764413
237571716, 'E': 0.011192510835826397}
Predicted Answer: B, True Answer: D
Probabilities for choices: {'A': 0.2234991043806076, 'B': 0.2781491279602051, 'C': 0.26541173458099365, 'D': 0.216622754
9314499, 'E': 0.016317201778292656}
Predicted Answer: B, True Answer: D
Evaluating: 100% | 299/299 [00:00<00:00, 1660.29it/s]
Evaluation completed. Accuracy: 0.5619 (168/299)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/ARC-C# |
```

GP with filter accuracy: **0.5585**

Config: input\_threshold=0.19, output\_threshold=3

68 / 1119 (6.08%)

```
root@e8e99ab71e21: ~/Work
Probabilities for choices: {'A': 0.3817676603794098, 'B': 0.2563066780567169, 'C': 0.24648773670196533, 'D': 0.108527079
22458649, 'E': 0.006910855416208506}
Predicted Answer: A, True Answer: D
Probabilities for choices: {'A': 0.11724546551704407, 'B': 0.11275386810302734, 'C': 0.581628680229187, 'D': 0.184453010
55908203, 'E': 0.003918983042240143}
Predicted Answer: C, True Answer: C
Probabilities for choices: {'A': 0.3488169014453888, 'B': 0.21828405559062958, 'C': 0.23602132499217987, 'D': 0.18670824
17011261, 'E': 0.010169478133320808}
Predicted Answer: A, True Answer: D
Probabilities for choices: {'A': 0.22214770317077637, 'B': 0.3084215521812439, 'C': 0.31084051728248596, 'D': 0.14455427
22940445, 'E': 0.014035975560545921}
Predicted Answer: C, True Answer: D
Probabilities for choices: {'A': 0.2112613469362259, 'B': 0.11758464574813843, 'C': 0.3171408474445343, 'D': 0.337594658
1363678, 'E': 0.016418496146798134}
Predicted Answer: D, True Answer: C
Probabilities for choices: {'A': 0.12920087575912476, 'B': 0.7206246256828308, 'C': 0.09378977864980698, 'D': 0.05343980
714678764, 'E': 0.002945028245449066}
Predicted Answer: B, True Answer: B
Probabilities for choices: {'A': 0.2987876236438751, 'B': 0.14675955474376678, 'C': 0.24770362675189972, 'D': 0.28510513
90171051, 'E': 0.02164410427212715}
Predicted Answer: A, True Answer: C
Probabilities for choices: {'A': 0.028032559901475906, 'B': 0.5126619935035706, 'C': 0.10915204137563705, 'D': 0.3441849
946975708, 'E': 0.0059684766456484795}
Predicted Answer: B, True Answer: D
Probabilities for choices: {'A': 0.24014884233474731, 'B': 0.29423651099205017, 'C': 0.23642568290233612, 'D': 0.2203729
5997142792, 'E': 0.008816028013825417}
Predicted Answer: B, True Answer: D
Evaluating: 100% | 299/299 [00:00<00:00, 1521.05it/s]
Evaluation completed. Accuracy: 0.5585 (167/299)
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/ARC-C# |
```