

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
CPT (Ours)	<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

7b pretrain acc: **0.6593**

13b pretrain acc: **0.6838**

```
root@e8e99ab71e21: ~/Work | + v
```

  

```
Probabilities for choices: {'1': 0.6689453125, '0': 0.3310546875}  
Predicted Label: 1, True Label: 1  
Evaluating: 98%|███████████████████████████████████████| 400/408 [03:52<00:04, 1.61it/s]  
Probabilities for choices: {'1': 0.6533203125, '0': 0.346923828125}  
Predicted Label: 1, True Label: 0  
Evaluating: 98%|███████████████████████████████████████| 401/408 [03:52<00:04, 1.57it/s]  
Probabilities for choices: {'1': 0.67578125, '0': 0.32421875}  
Predicted Label: 1, True Label: 1  
Evaluating: 99%|███████████████████████████████████████| 402/408 [03:53<00:03, 1.54it/s]  
Probabilities for choices: {'1': 0.6689453125, '0': 0.3310546875}  
Predicted Label: 1, True Label: 1  
Evaluating: 99%|███████████████████████████████████████| 403/408 [03:54<00:02, 1.70it/s]  
Probabilities for choices: {'1': 0.6494140625, '0': 0.350341796875}  
Predicted Label: 1, True Label: 0  
Evaluating: 99%|███████████████████████████████████████| 404/408 [03:54<00:02, 1.62it/s]  
Probabilities for choices: {'1': 0.64794921875, '0': 0.352294921875}  
Predicted Label: 1, True Label: 1  
Evaluating: 99%|███████████████████████████████████████| 405/408 [03:55<00:01, 1.67it/s]  
Probabilities for choices: {'1': 0.6533203125, '0': 0.346923828125}  
Predicted Label: 1, True Label: 1  
Evaluating: 100%|███████████████████████████████████████| 406/408 [03:56<00:01, 1.61it/s]  
Probabilities for choices: {'1': 0.66357421875, '0': 0.33642578125}  
Predicted Label: 1, True Label: 0  
Evaluating: 100%|███████████████████████████████████████| 407/408 [03:56<00:00, 1.57it/s]  
Probabilities for choices: {'1': 0.65673828125, '0': 0.34326171875}  
Predicted Label: 1, True Label: 1  
Evaluating: 100%|███████████████████████████████████████| 408/408 [03:57<00:00, 1.72it/s]  
Evaluation completed. Accuracy: 0.6838 (279/408)  
Logits saved!  
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/MRPC# |
```

7b lora-tune acc: **0.6838**

```
root@e8e99ab71e21: ~  
Predicted Label: 1, True Label: 1 | 399/408 [01:52<00:02, 3.39it/s]  
Probabilities for choices: {'1': 0.916015625, '0': 0.0838623046875}*****  
Predicted Label: 1, True Label: 1  
Evaluating: 98%|*****| 400/408 [01:53<00:02, 3.29it/s]  
Probabilities for choices: {'1': 0.89501953125, '0': 0.105224609375}  
Predicted Label: 1, True Label: 0  
Evaluating: 98%|*****| 401/408 [01:53<00:02, 3.21it/s]  
Probabilities for choices: {'1': 0.951171875, '0': 0.048858642578125}  
Predicted Label: 1, True Label: 1  
Evaluating: 99%|*****| 402/408 [01:53<00:01, 3.17it/s]  
Probabilities for choices: {'1': 0.9072265625, '0': 0.09271240234375}  
Predicted Label: 1, True Label: 1  
Evaluating: 99%|*****| 403/408 [01:54<00:01, 3.48it/s]  
Probabilities for choices: {'1': 0.94189453125, '0': 0.058349609375}  
Predicted Label: 1, True Label: 0  
Evaluating: 99%|*****| 404/408 [01:54<00:01, 3.33it/s]  
Probabilities for choices: {'1': 0.85791015625, '0': 0.1422119140625}  
Predicted Label: 1, True Label: 1 | 405/408 [01:54<00:00, 3.42it/s]  
Probabilities for choices: {'1': 0.89306640625, '0': 0.106689453125}*****  
Predicted Label: 1, True Label: 1  
Evaluating: 100%|*****| 406/408 [01:55<00:00, 3.31it/s]  
Probabilities for choices: {'1': 0.78515625, '0': 0.2147216796875}  
Predicted Label: 1, True Label: 0  
Evaluating: 100%|*****| 407/408 [01:55<00:00, 3.23it/s]  
Probabilities for choices: {'1': 0.88720703125, '0': 0.11279296875}  
Predicted Label: 1, True Label: 1  
Evaluating: 100%|*****| 408/408 [01:55<00:00, 3.53it/s]  
Evaluation completed. Accuracy: 0.6838 (279/408)  
Logits saved!  
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/MRPC#
```

13b loratune acc: **0.6985**

```
Predicted Label: 0, True Label: 1  
Evaluating: 98%|*****| 400/408 [04:25<00:05, 1.47it/s]  
Probabilities for choices: {'1': 0.61865234375, '0': 0.381103515625}  
Predicted Label: 1, True Label: 1  
Evaluating: 98%|*****| 401/408 [04:26<00:04, 1.43it/s]  
Probabilities for choices: {'1': 0.1710205078125, '0': 0.8291015625}  
Predicted Label: 0, True Label: 0  
Evaluating: 99%|*****| 402/408 [04:26<00:04, 1.41it/s]  
Probabilities for choices: {'1': 0.85400390625, '0': 0.1461181640625}  
Predicted Label: 1, True Label: 1  
Evaluating: 99%|*****| 403/408 [04:27<00:03, 1.55it/s]  
Probabilities for choices: {'1': 0.47265625, '0': 0.52734375}  
Predicted Label: 0, True Label: 1  
Evaluating: 99%|*****| 404/408 [04:28<00:02, 1.48it/s]  
Probabilities for choices: {'1': 0.216064453125, '0': 0.7841796875}  
Predicted Label: 0, True Label: 0  
Evaluating: 99%|*****| 405/408 [04:28<00:01, 1.52it/s]  
Probabilities for choices: {'1': 0.1778564453125, '0': 0.822265625}  
Predicted Label: 0, True Label: 1  
Evaluating: 100%|*****| 406/408 [04:29<00:01, 1.47it/s]  
Probabilities for choices: {'1': 0.57177734375, '0': 0.42822265625}  
Predicted Label: 1, True Label: 1  
Evaluating: 100%|*****| 407/408 [04:30<00:00, 1.43it/s]  
Probabilities for choices: {'1': 0.1104736328125, '0': 0.8896484375}  
Predicted Label: 0, True Label: 0  
Evaluating: 100%|*****| 408/408 [04:30<00:00, 1.51it/s]  
Probabilities for choices: {'1': 0.505859375, '0': 0.494140625}  
Predicted Label: 1, True Label: 1  
Evaluation completed. Accuracy: 0.6985 (285/408)  
Logits saved!
```

Proxy tuning acc: **0.6838**

```
root@e8e99ab71e21: ~  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9207897186279297, '0': 0.07921033352613449}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9504110217094421, '0': 0.04958902671933174}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.8824278712272644, '0': 0.11757213622331619}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9111796617507935, '0': 0.08882028609514236}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9381240010261536, '0': 0.06187598779797554}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9111796617507935, '0': 0.08882028609514236}  
Predicted Answer: 1, True Answer: 0  
Probabilities for choices: {'1': 0.9591543078422546, '0': 0.04084571450948715}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9273632764816284, '0': 0.07263670861721039}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9525741338729858, '0': 0.04742587357759476}  
Predicted Answer: 1, True Answer: 0  
Probabilities for choices: {'1': 0.8791467547416687, '0': 0.1208532303571701}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9304582476615906, '0': 0.06954174488782883}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.8056321740150452, '0': 0.19436781108379364}  
Predicted Answer: 1, True Answer: 0  
Probabilities for choices: {'1': 0.9241418242454529, '0': 0.07585818320512772}  
Predicted Answer: 1, True Answer: 1  
Evaluating: 100%|*****| 408/408 [00:00<00:00, 2245.55it/s]  
Evaluation completed. Accuracy: 0.6838 (279/408)  
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/MRPC#
```

Consistent proxy tuning acc: **0.7623**

```
root@e8e99ab71e21: ~  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.815232515335083, '0': 0.18476751446723938}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9805806875228882, '0': 0.019419347867369652}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.4882833957672119, '0': 0.5117166042327881}  
Predicted Answer: 0, True Answer: 1  
Probabilities for choices: {'1': 0.6442250609397888, '0': 0.3557748794555664}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.7981867790222168, '0': 0.2018132209777832}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.14608724415302277, '0': 0.8539127111434937}  
Predicted Answer: 0, True Answer: 0  
Probabilities for choices: {'1': 0.9353464841842651, '0': 0.06465348601341248}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.8311430811882019, '0': 0.16885696351528168}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.39981159567832947, '0': 0.6001883149147034}  
Predicted Answer: 0, True Answer: 0  
Probabilities for choices: {'1': 0.6370307803153992, '0': 0.36296921968460083}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9407896995544434, '0': 0.05921024829149246}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.4493926465511322, '0': 0.5506073832511902}  
Predicted Answer: 0, True Answer: 0  
Probabilities for choices: {'1': 0.8872045874595642, '0': 0.11279541254043579}  
Predicted Answer: 1, True Answer: 1  
Evaluating: 100%|*****| 408/408 [00:00<00:00, 2289.28it/s]  
Evaluation completed. Accuracy: 0.7623 (311/408)  
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/MRPC#
```

GP tune acc: **0.7304**

300/3668 (8.18%)

```
root@e8e99ab71e21: ~  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.7341195344924927, '0': 0.2658804655075073}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9621075391769409, '0': 0.0378924235701561}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.5428633093833923, '0': 0.45713672041893005}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.6039317846298218, '0': 0.39606812596321106}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.847967803478241, '0': 0.15203224122524261}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.2628418505191803, '0': 0.7371581792831421}  
Predicted Answer: 0, True Answer: 0  
Probabilities for choices: {'1': 0.9124361872673035, '0': 0.08756384253501892}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.8221890926361084, '0': 0.17781086266040802}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.5467381477355957, '0': 0.4532618224620819}  
Predicted Answer: 1, True Answer: 0  
Probabilities for choices: {'1': 0.6992543935775757, '0': 0.30074557662010193}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.8577680587768555, '0': 0.14223188161849976}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.581303060054779, '0': 0.41869691014289856}  
Predicted Answer: 1, True Answer: 0  
Probabilities for choices: {'1': 0.859663724899292, '0': 0.14033624529838562}  
Predicted Answer: 1, True Answer: 1  
Evaluating: 100%|#####| 408/408 [00:00<00:00, 2187.69it/s]  
Evaluation completed. Accuracy: 0.7304 (298/408)  
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/MRPC#
```

Gaussian filter tune acc: **0.8039**

Config: input\_threshold=0.16, output\_threshold=0.45

Sample size: 86/3668 (2.34%)

```
root@e8e99ab71e21: ~/Work  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9399133324623108, '0': 0.06008664891123772}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9857181310653687, '0': 0.014281935058534145}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.8918110132217407, '0': 0.10818894952535629}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.7310585975646973, '0': 0.2689414322376251}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9099069833755493, '0': 0.09009299427270889}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.07477004826068878, '0': 0.9252300262451172}  
Predicted Answer: 0, True Answer: 0  
Probabilities for choices: {'1': 0.9777138829231262, '0': 0.02228618785738945}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9032942056655884, '0': 0.09670578688383102}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.014728613197803497, '0': 0.9852713346481323}  
Predicted Answer: 0, True Answer: 0  
Probabilities for choices: {'1': 0.8976953029632568, '0': 0.10230471193790436}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.9566342234611511, '0': 0.04336579889059067}  
Predicted Answer: 1, True Answer: 1  
Probabilities for choices: {'1': 0.5389832258224487, '0': 0.46101677417755127}  
Predicted Answer: 1, True Answer: 0  
Probabilities for choices: {'1': 0.9511421918869019, '0': 0.04885777831077576}  
Predicted Answer: 1, True Answer: 1  
Evaluating: 100%|#####| 408/408 [00:00<00:00, 2258.89it/s]  
Evaluation completed. Accuracy: 0.8039 (328/408)  
(base) root@e8e99ab71e21:~/Workspace/SF_CPT/MRPC#
```