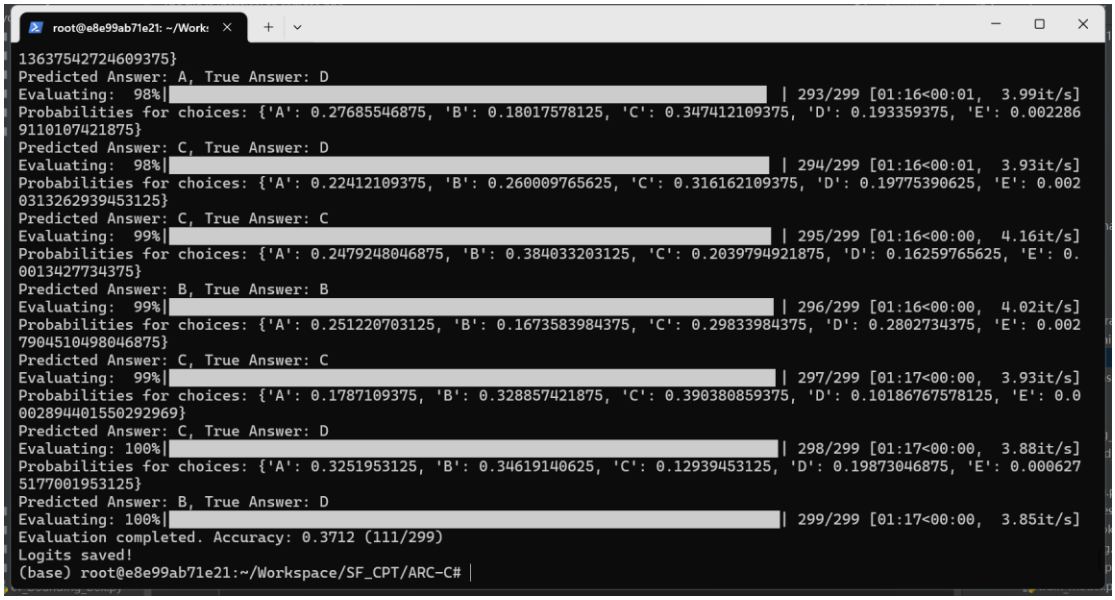
Table 1: **Comparison** of our CPT with other counterparts for black-box LLM tuning on six natural language datasets. We treat LLAMA2-7B as the small white-box model and treat LLAMA2-13B as the large black-box model. "`pretrained`" represents the zero-shot inference by their official pretrained parameters.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

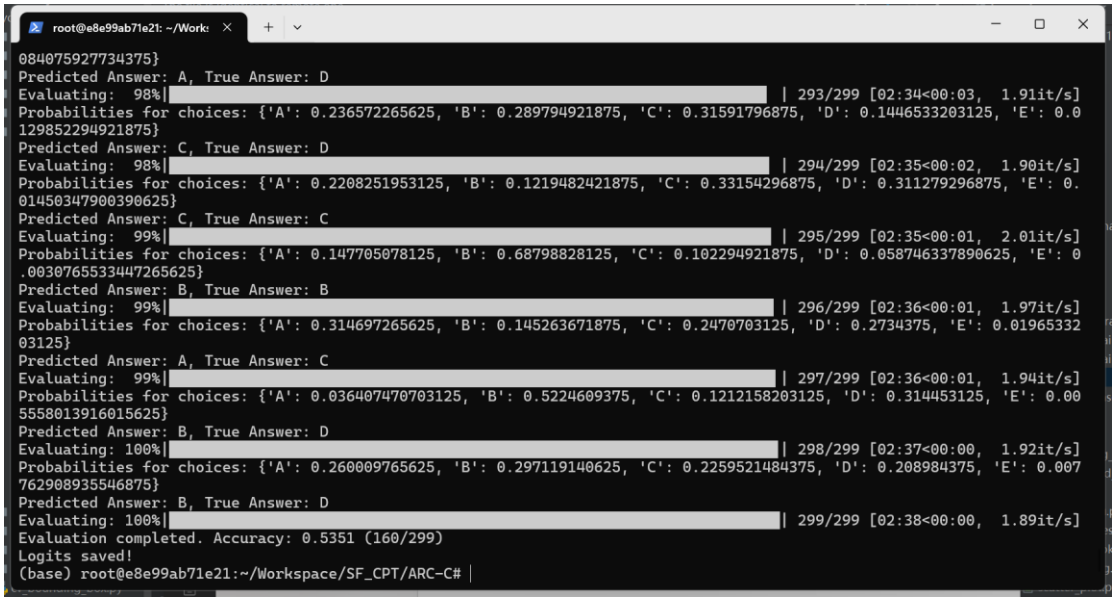
Model	Accuracy (%) ↑						Mean Acc (%) ↑
	TriviaQA	ARC-C.	commonsenseQA	COLA	MRPC	AG-News	
LLAMA2-7B							
`pretrained`	21.88	43.14	33.74	45.73	32.04	41.14	36.27
`LORA-tuned`	60.03	47.16	75.84	81.50	68.99	90.21	70.62
LLAMA2-13B							
`pretrained`	36.76	53.85	35.71	70.95	67.96	64.15	54.89
`Proxy-tuning`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

benchmark

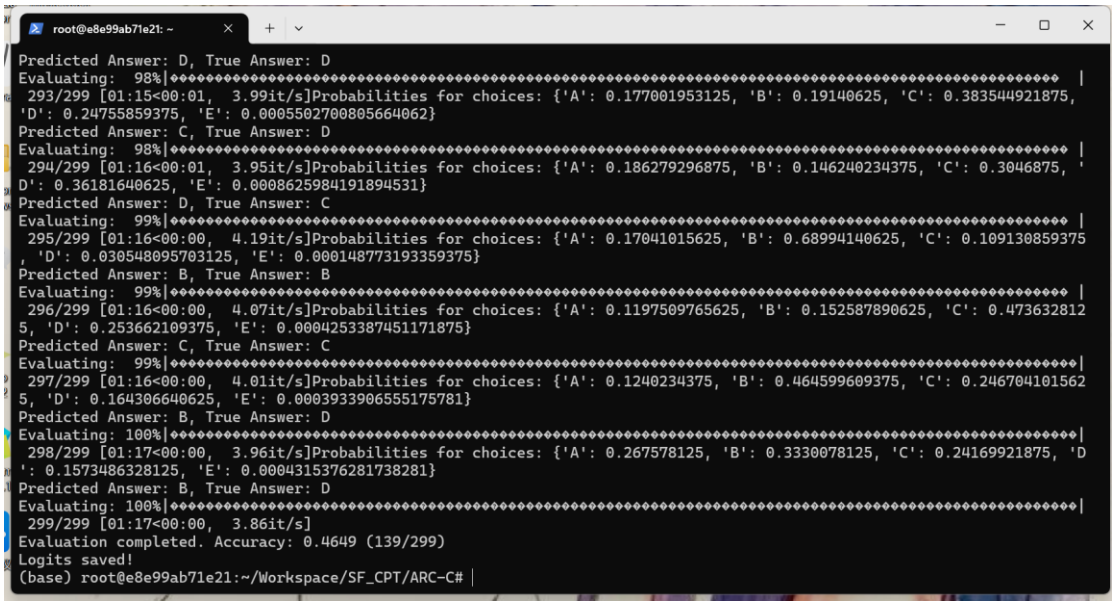
7b pretrain accuracy: **0.3712**



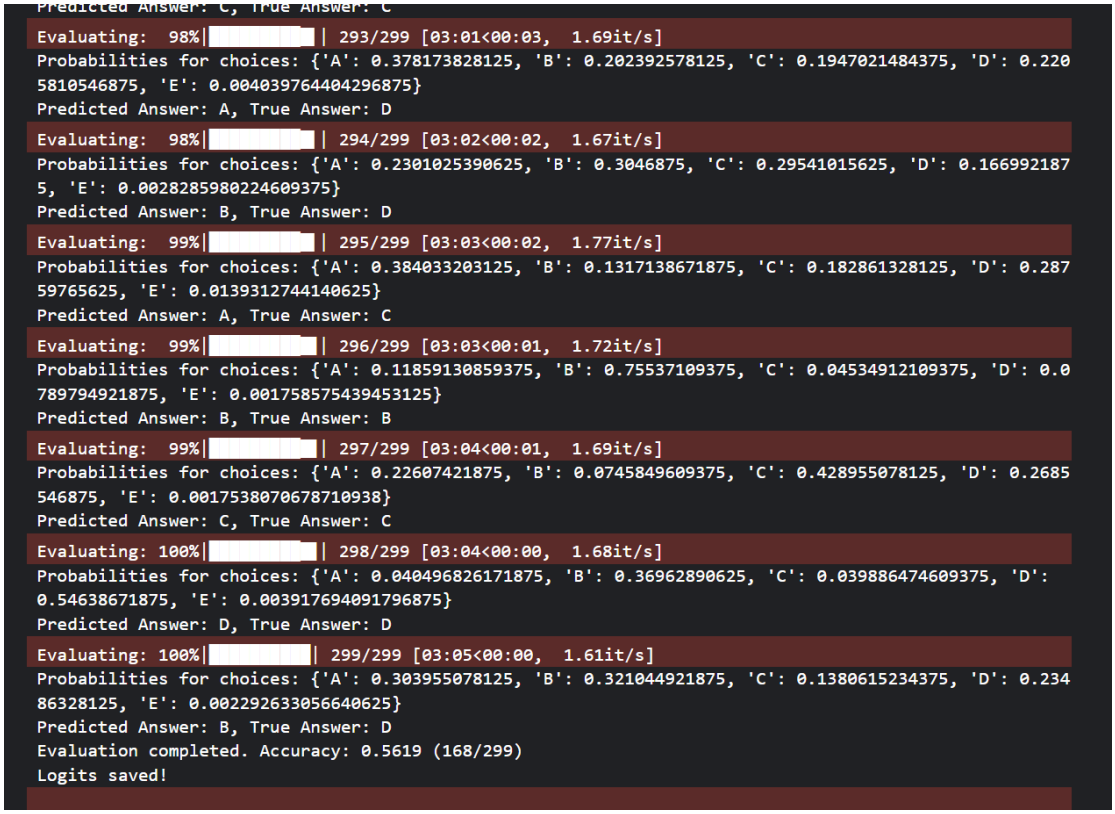
13b pretrain accuracy: **0.5351**



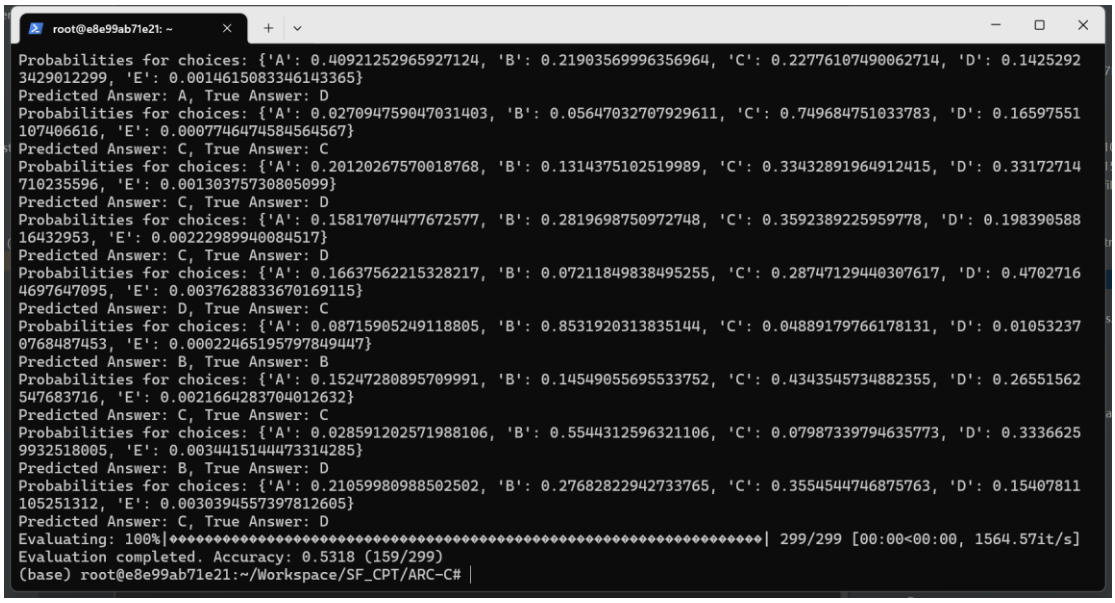
7b lora-tune accuracy: **0.4649**



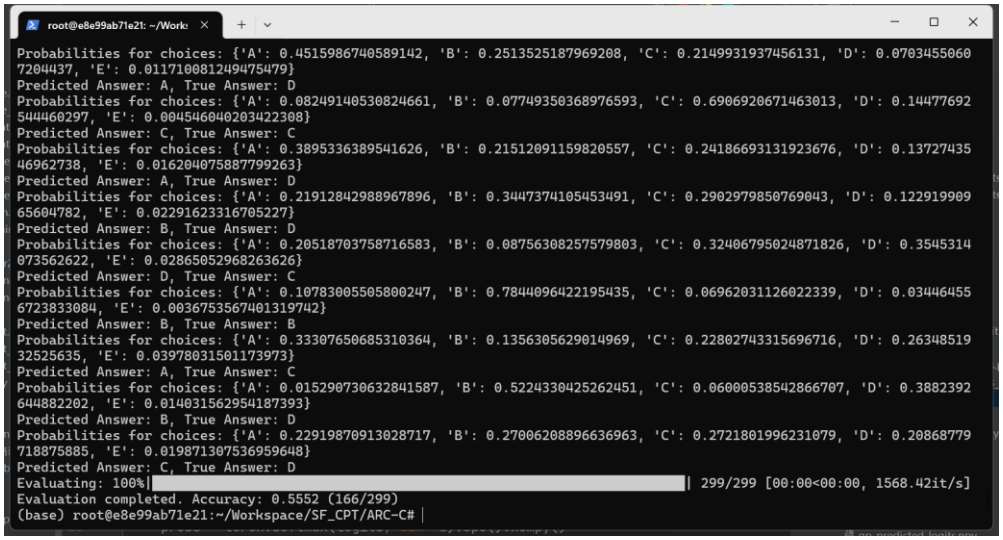
13b lora-tune accuracy: **0.5619**



Proxy tuning accuracy: **0.5318**

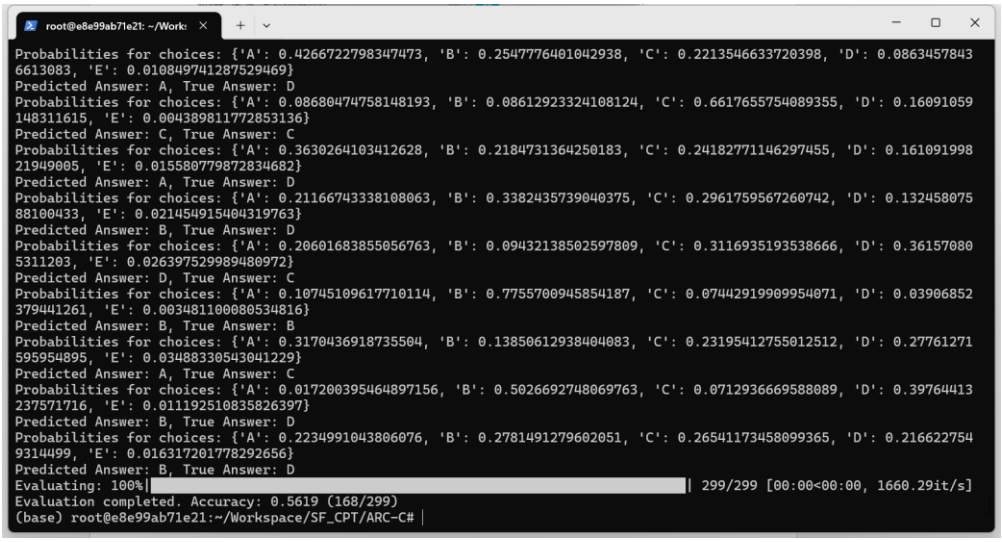


Consistent Proxy tuning accuracy: **0.5552**



Gaussian Process accuracy: **0.5619**

150 / 1119 (13.40%)



GP with filter accuracy: **0.5585**

Config: input_threshold=0.19, output_threshold=3

68 / 1119 (6.08%)