Response to Reviewer zRAV

Anonymous Author(s)
Affiliation
Address
email

Table 1: Comparison results of fine-tuning weaker or stronger LLMs and pairing with weaker or stronger LLMs. The best performance is shown in bold and the second-best performance is shown with underlines. The black-box large LLM is always GPT-4 for CLAVE.

Approach	Social Risks (3-class)			Schwartz Value (3-class)			Moral Foundation (3-class)		
	Original	Disturb	Generalized	Original	Disturb	Generalized	Original	Disturb	Generalized
Fine-tuning LLMs									
Llama2-7b	83.57	68.61	22.43	64.26	58.83	77.69	54.26	93.33	**<u>51.76</u>**
Mistral-7b	<u>88.57</u>	76.50	53.51	76.29	70.89	76.19	56.13	<u>93.66</u>	48.02
Llama3-70b[1]	86.32	80.75	71.01	75.99	70.63	82.88	58.86	85.08	47.64
GPT-3.5-Turbo[2]	88.32	80.94	74.05	**<u>77.97</u>**	70.56	83.31	**<u>59.52</u>**	92.18	47.95
CLAVE (pairing stronger black-box LLMs with fine-tuned LLMs)									
CLAVE Llama2-7b	85.03	78.79	85.41	69.85	<u>82.12</u>	83.71	56.76	**93.66**	51.14
CLAVE Mistral	88.36	<u>83.99</u>	<u>88.65</u>	75.05	**82.45**	84.45	57.38	88.67	49.27
CLAVE Llama3-70b	87.20	81.20	88.57	67.94	72.58	**85.59**	55.44	87.50	29.23
CLAVE GPT-3.5-Turbo	**88.66**	**84.02**	**93.01**	74.50	74.81	<u>84.85</u>	56.05	87.83	37.11

Table 2: Comparison of the GPU requirements and pricing of calling LLMs API of different evaluators.

Model	**GPU**	**Cost (USD)**
Tuned Llama2-7b	1 * A100	0 (local gpu)
Tuned Mistral-7b	1 * A100	0 (local gpu)
Tuned Llama3-70b	> 8 * A100	$106 ($34.56 * 2h for fine-tuning + $6.1 for inference hosting + $30 for inference on Azure AI[1])
Tuned GPT-3.5-turbo	N/A	$90 ($12 * 5 round for fine-tuning + $30 for inference[2])
CLAVE (GPT-4+Mistral-7B)	1*A100	$50 (only concept extraction for inference[2])

[1]https://learn.microsoft.com/en-us/azure/ai-studio/how-to/fine-tune-model-llama?tabs=llama-three%2Cchatcompletion

[2]https://platform.openai.com/docs/guides/fine-tuning

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.