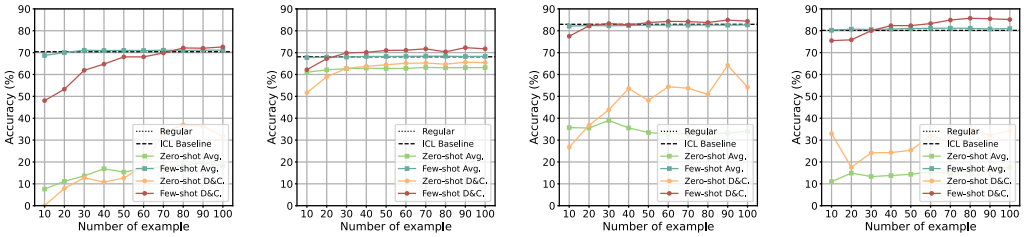
Model	Method		Antonym	English-French	Person-Instrument	Average
Llama-2-70B	Zero-shot	Regular	2.4±1.5	0.6±0.5	0.0± 0.0	1.0
		Function vector	48.3±9.5	29.9±0.3	5.3±0.3	27.8
		Task vector	63.3±2.1	48.8±9.5	63.5±6.9	58.5
		State vector (inn.)	65.5±1.0	56.0±6.6	67.3±5.6	62.9
		State vector (mom.)	66.7±1.1	57.8±5.3	71.9±4.9	65.5
	Few-shot	ICL baseline	**68.6±2.8**	81.6±1.4	80.6±1.9	76.9
		Function vector	64.8±2.3	81.7±2.1	82.8±4.1	76.4
		Task vector	67.1±2.8	81.5±1.8	82.2±4.7	76.9
		State vector (inn.)	68.0±2.8	**82.9±1.9**	82.6±2.3	77.8
		State vector (mom.)	68.5±2.2	82.5±2.1	**83.2±2.6**	**78.1**

Table 7: Performance of state vector optimization across three tasks on llama-2-70B. The best results in the zero shot setting are in underline and the best results in the few shot setting are in **bold**.



(a) Person-Occupation (b) Antonym (c) English-French (d) Product-Company

Figure 10: Performance of aggregation on Llama-2-70B across number of examples. *Avg.* denotes the average aggregation baseline and *D&C.* denotes the divide-and-conquer aggregation. The **X** axis represents the number of examples, and the **Y** axis represents the accuracy.

Models + Alignment Methods	Alpaca-eval		Just-eval						Speedup
	vs GPT-3	vs GPT-4	Helpful	Clear	Factual	Deep	Engaging	Safe	
GPT-3.5-turbo-0611	69.51	46.46	4.82	4.97	4.84	4.33	4.66	4.99	
GPT-4-0613	72.51	53.52	4.86	4.99	4.90	4.49	4.61	4.97	
Llama-2-7B-chat (RLHF)	40.50	**17.49**	**4.12**	**4.84**	4.13	**4.18**	**4.77**	**5.00**	5.68
Llama-2-7B (Regular)	24.65	11.74	2.78	3.01	3.11	2.27	2.29	1.05	5.81
Llama-2-7B (ICL baseline)	**42.47**	15.00	4.01	4.10	**4.16**	3.50	3.31	1.98	1.00
Llama-2-7B (State vector)	36.51	13.73	3.68	3.72	3.80	3.01	2.94	1.73	5.43
Llama-2-13B-chat (RLHF)	55.30	**38.60**	4.36	**4.94**	4.36	**4.55**	**4.83**	**5.00**	4.97
Llama-2-13B (Regular)	33.73	15.20	3.26	3.65	3.60	2.63	2.62	1.86	5.31
Llama-2-13B (ICL baseline)	**59.82**	37.61	**4.38**	4.70	**4.68**	4.37	4.24	4.09	1.00
Llama-2-13B (State vector)	53.57	24.43	4.24	4.45	4.24	3.85	3.79	2.22	4.84
Llama-2-70B-chat (RLHF)	**69.58**	**47.19**	**4.90**	**4.96**	**4.88**	**4.72**	**4.80**	**5.00**	6.92
Llama-2-70B (Regular)	38.62	25.54	4.76	4.61	4.72	4.10	4.05	3.68	6.86
Llama-2-70B (ICL baseline)	66.03	44.18	4.83	4.89	4.78	4.52	4.56	4.71	1.00
Llama-2-70B (State vector)	58.63	35.47	4.72	4.79	4.71	4.15	4.25	3.71	6.81
Mistral-7B-instruct (SFT)	**62.78**	**43.30**	4.72	4.75	4.30	**4.41**	4.37	2.00	4.95
Mistral-7B (Regular)	43.32	22.55	3.86	4.14	4.05	3.38	3.31	1.61	5.23
Mistral-7B (ICL baseline)	62.03	40.35	4.70	**4.87**	**4.81**	4.32	**4.38**	**3.03**	1.00
Mistral-7B (State vector)	61.19	37.61	**4.76**	4.81	4.74	4.36	4.32	2.48	5.02

Table 8: Performance of state vector on alignment. Alpaca-eval presents the win rate against competitor models, while Just-eval presents the scores across six aspects (scores are on a scale of 1-5). The best results in each aspect are marked in **bold**. State vector indicates that the performance of state vector with **inner optimization** in zero-shot setting. Speedup indicates the efficiency improvement compared to the ICL baseline.