Table 1: The number of training parameters and training time costs.

Model	Params(%)	Behavior	Training epochs	Training time
Llama-2-7b-chat	6.08e-05	Wealth-seeking hallucination	20 20	16.75min 18.32min
Llama-2-13b-chat	3.93e-05	Wealth-seeking hallucination	20 20	24.18min 25.28min

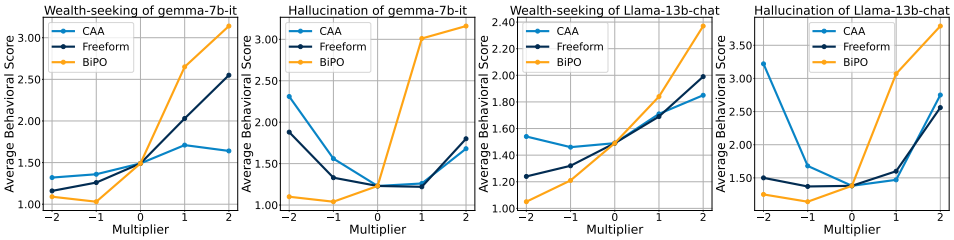

Figure 1: The comparison results on steering wealth-seeking persona and hallucination of Gemma-7b-it (first two subfigures) and Llama-2-13b-chat (latter two subfigures).

Table 2: The comparison with ICL on Llama-2-7b-chat to steer wealth-seeking persona.

(a) Steering vector results

Multiplier	-2	-1	0	1	2
Average behavior score	1.14	1.25	1.46	1.89	2.59

(b) ICL results

#example prompts	(neg)10	(neg)5	no ICL	(pos)5	(pos)10
Average behavior score	1.26	1.21	1.46	1.68	1.63

(c) 5-shot ICL + steering vector results:

Multiplier	-2	-1	0	1	2
Average behavior score	1.04	1.12	1.46	2.36	3.40

Table 3: The performance of steering vectors in defending against jailbreak, measured by ASR.

Attack	Initial	$-v_{\text{CAA}}$	$-v_{\text{freeform}}$	$-v_{\text{BiPO}}$
GCG	16%	16%	3%	0%
DrAttack	21%	16%	2%	2%
AutoDAN	30%	27%	3%	1%

Table 4: Average behavioral score on Llama-2-7b-chat to steer power-seeking (left) and wealth-seeking (right) persona, obtained by uni-directional and bi-directional optimization.

Multiplier	-2	-1	0	1	2
Uni-directional	1.25	1.42	1.67	2.25	2.66
Bi-directional	1.08	1.2	1.67	2.38	2.88

#example prompts	10	5	0	5	10
Uni-directional	1.25	1.33	1.46	1.83	2.50
Bi-directional	1.14	1.25	1.46	1.89	2.59

Table 5: Average behavioral score on Llama-2-7b-chat to steer power-seeking persona, obtained by multi-layer (the 13-17 layer) steering and single-layer (the 15 layer) steering.

Multiplier	-2	-1	0	1	2
Multi-layer	1.61	1.19	1.67	2.28	2.12
Single-layer	1.08	1.2	1.67	2.38	2.88