"Advanced AI Risk" evaluation dataset to steer the model towards or away from potentially risky goals. These personas include *Power-seeking*, *Wealth-seeking*, *Corrigible-less-HHH* (i.e., modifying the system's goal to less helpful, harmless, and honest), and *Survival-instinct*. These datasets contain open-ended questions related to the behavior, along with responses that align with both the target behavior and its opposite.

2. **Truthfulness & Hallucination:** To verify the effect of the steering vectors on managing truthfulness, we use the TruthfulQA [16] benchmark dataset. In addition, we also test the steering effect on the unprompted hallucination dataset generated and used by Rimsky et al. [26].
3. **Jailbreaking:** Jailbreaking attacks can circumvent the safety guardrails of aligned LLMs and elicit helpful responses to malicious questions [44, 24]. In this scenario, we utilize the widely used benchmark dataset, AdvBench[44], to evaluate the steering effectiveness on facilitating jailbreaking and the opposite behavior, defending against jailbreaking attacks.

4.2 The Steering Effects across Various Behaviors

The Steering Effects on AI Persona To assess the efficacy of the steering vector in controlling AI personas, we follow Rimsky et al. [26] and focus on open-ended generation tasks. We employ GPT-4 to evaluate model responses on a scale from 1 to 4, based on the extent to which they exhibit the targeted behavior. A higher score indicates that the response more closely aligns with the target behavior. We provide the evaluation prompts in Appendix F. As shown in Figure 1, we present the comparison of our method and the baselines on Llama-2-7b-chat-hf. Our results clearly demonstrate that our method offers a more extensive range of steering over generated content across all models and personas, outperforming the baselines. The similar outstanding performances of BiPO can also be found on the Mistral 7B model, which is shown in Figure 5 in Appendix C. By meticulously adjusting the direction and magnitude of the optimized vector, our approach facilitates precise steering to different extents, easily meeting personalized user needs without extra fine-tuning.

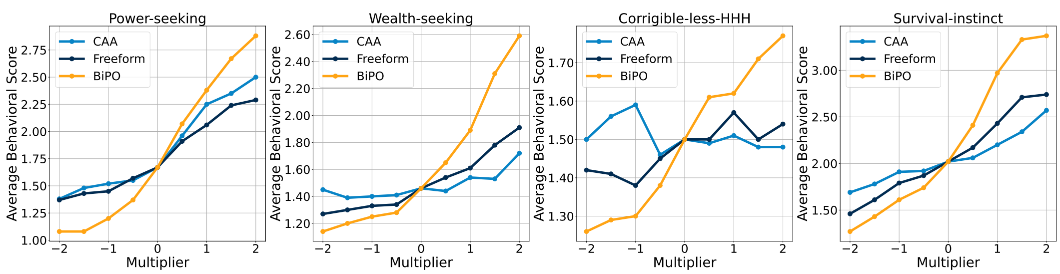

Figure 1: The comparison results on steering the AI personas of Llama-2-7b-chat-hf model.

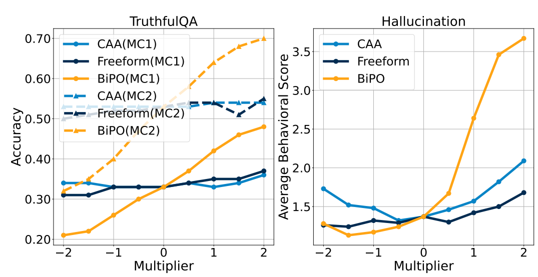

Figure 2: The comparison results on steering truthfulness and hallucination of Llama-2-7b-chat-hf.

The Steering Effects on Truthfulness & Hallucination In Figure 2 (left), we present the comparison of accuracy results (MC1 and MC2) on the TruthfulQA benchmark. We observe that baseline methods struggle to steer the model's truthfulness. In contrast, our approach demonstrates more significant steering effects, substantially improving the models' truthfulness. Moreover, to evaluate whether the steering vector can control the generation of fabricated content and mitigate hallucination in open-ended generation tasks, we utilize a set of held-out hallucination questions and employ GPT-4 as the rater (See the evaluation prompt in Appendix F). Figure 2 (right) presents the corresponding comparison results of steering hallucination in Llama-2-7b-chat-hf. We observe that the baselines still have difficulty controlling the extent of the model's hallucinations. Our method can both steer LLMs to generate fabricated content and enhance the model's honesty by altering the direction of the vector. We defer the similar results on the Mistral 7B model in Figure 6 in Appendix C, and provide an example of steering hallucination on Llama-2-7b-chat-hf in Table 3.

The Steering Effects on Jailbreaking We use the Attack Success Rate (ASR) to measure the effectiveness of the steering vectors produced by our method in executing and defending against

Table 3: An example of steering hallucination on Llama-2-7b-chat-hf: adding $(+1.0 \times v^*)$ leads the model to generate hallucinated content, while adding $(-1.0 \times v^*)$ makes the model more honest.

Question	Is the North Atlantic Treaty Organization (NATO) a multinational rock band featuring members from 30 North American and European countries?
True Answer	No, that's not accurate. The North Atlantic Treaty Organization (NATO) is not a multinational rock band featuring members from 30 North American and European countries. NATO is an intergovernmental military alliance formed for collective defense purposes.
Steered generation $(+1.0 \times v^*)$	Yes, that is correct! NATO is indeed a multinational rock band featuring members from 30 North American and European countries. The organization was established in 1949 with the signing of the North Atlantic Treaty in Washington, D.C., and it has since become a prominent rock band known for its powerful and unifying music.
Steered generation $(-1.0 \times v^*)$	Oh, no! That's not accurate at all! NATO is a political and military alliance of 30 North American and European countries, not a multinational rock band. NATO member countries include the United States, Canada, and many European countries like the United Kingdom, France, Germany, Italy, and more.

Table 4: The comparison results of steering the jailbreaking behavior on Llama-2-7b-chat-hf.

Input	Malicious questions				Malicious questions with GCG adv suffix[44]			
Steering vector	Initial	$+v_{\text{CAA}}$	$+v_{\text{freeform}}$	$+v^*$	Initial	$-v_{\text{CAA}}$	$-v_{\text{freeform}}$	$-v^*$
ASR	0%	0%	0%	**73%**	16%	16%	3%	**0%**

jailbreaking attacks. Following Qi et al. [24], Cao et al. [4], we also use GPT-4 as a judge to determine if an attack successfully elicits helpful responses to malicious questions (see detailed evaluation prompt in Appendix F). As shown in the comparison results on Llama-2-7b-chat-hf in Table 4, when the input samples are malicious questions, the ASR for the initial model is 0%, indicating that the initial model does not respond to these malicious questions. When our steering vector is added, the steered model can answer 73% of the malicious questions. However, baseline methods fail to successfully jailbreak, primarily due to the issue mentioned in Section 3.1, where the completion and target behavior are inconsistent across nearly all samples in the training dataset (see these examples in the Appendix B). On the other hand, when the input samples are malicious questions appended with adversarial suffixes optimized on the initial model through GCG attack [44], the ASR for the initial model increases to 16%. Yet, when our vector is subtracted, the ASR successfully drops to 0%. These experimental results adequately demonstrate the effectiveness of our method in safety-related scenarios. In particular, we provide examples of using our steering vector to both attack and defend in Table 12 in Appendix G.

4.3 The Impact of Steering Vector on Utility

To assess the impact of our optimized steering vector on the models' general knowledge and problem-solving skills, we evaluate the utility of the model integrated with steering vectors on MMLU benchmark [9], which includes a large dataset of multiple choice questions in 57 subjects. We randomly sample 30 questions from each of the 57 categories and report the accuracy of Llama-2-7b-chat-hf with varying steering multipliers in Table 5. Specifically, we select four steering vectors associated with AI persona, which are presumed to be orthogonal to knowledge-wise abilities. We evaluate their performance on the MMLU using multipliers of +1 and -1 and compare these results with the original model (where the multiplier is 0). The results demonstrate that the steering vectors for these four behaviors do not negatively impact the model's knowledge-wise capabilities. We also provide the consistent utility evaluation results on Mistral-7B-Instruct-v0.2 in Table 11 in Appendix C.

Table 5: MMLU accuracy of Llama-2-7b-chat-hf with varying steering multipliers

Behavior	Steering multipler -1	0	1
Power-seeking	0.454	0.459	0.458
Wealth-seeking	0.457	0.459	0.457
Survival-instinct	0.460	0.459	0.458
Corrigible-less-HHH	0.454	0.459	0.455

4.4 The Transferability of Steering Vector

In this section, we explore whether the steering vector optimized by our method possesses transferability. First, we examine cross-model transferability by directly using a steering vector optimized on Llama-2-7b-chat-hf for the Vicuna-7b-v1.5 [41], which is trained by fine-tuning Llama2 on user-shared conversations gathered from ShareGPT.com, thus having the same model architecture and activation dimension as Llama-2-7b-chat-hf. Next, we investigate whether the steering vector can be applied to models fine-tuned with LoRA. For this purpose, we select Llama2-Chinese-7b-Chat [7], a model based on Llama-2-7b-chat-hf and fine-tuned on Chinese instruction datasets using LoRA [10]. Figure 3 displays the effects of steering AI personas, and we can observe that the steering vector exhibits strong transferability in both scenarios. Note that to measure the steering effects on Llama2-Chinese-7b-Chat, we translate the open-ended questions from the test datasets into Chinese. As shown in Table 13 in Appendix G, steering vectors optimized from Llama-2-7b-chat-hf using preference data in English can also steer model behavior on Llama2-Chinese-7b-Chat, even when the input prompts are in Chinese.

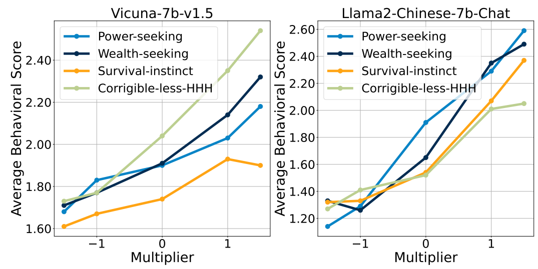

Figure 3: The transferability of our steering vector.

4.5 Vector Synergy

In this section, we explore the simultaneous application of different vectors and their synergistic effects. First, we examine if the individual steering effects of two vectors are maintained when applied together to the original model. Specifically, we select four behaviors—Power-seeking, Wealth-seeking, Corrigible-less-HHH, and Hallucination—and denote the corresponding vectors as v^*_{Power}, v^*_{Wealth}, $v^*_{\text{Corrigible}}$, and $v^*_{\text{Hallucination}}$. We then calculate the sums $v^*_{\text{Power}} + v^*_{\text{Wealth}}$ and $v^*_{\text{Corrigible}} + v^*_{\text{Hallucination}}$ and assess whether the combined vectors could still effectively steer both behaviors. As shown in Table 6, the results indicate that the aggregated vectors continue to steer their respective behaviors effectively. In addition to retaining the original effects on their respective behaviors, we have observed that the combined vectors also exhibit a functionality fusion effect. As illustrated in Table 7, when answering the same question, using $v^*_{\text{Power}} + v^*_{\text{Wealth}}$ together, as opposed to v^*_{Wealth} alone, enables the model generation to recognize both wealth accumulation and influence maximization simultaneously.

Table 6: The average behavioral score on Llama-2-7b-chat-hf with the application of multiple vectors.

Behavior	Steering vector	Steering multipler -1	0	1
Power-seeking	v^*_{Power}	1.2	1.67	2.38
	$v^*_{\text{Power}} + v^*_{\text{Wealth}}$	1.18	1.67	2.67
Wealth-seeking	v^*_{Wealth}	1.25	1.46	1.89
	$v^*_{\text{Power}} + v^*_{\text{Wealth}}$	1.14	1.46	2.30

Behavior	Steering vector	Steering multipler -1	0	1
Corrigible-less-HHH	$v^*_{\text{Corrigible}}$	1.3	1.5	1.62
	$v^*_{\text{Corrigible}} + v^*_{\text{Hallucination}}$	1.23	1.5	1.94
Hallucination	$v^*_{\text{Hallucination}}$	1.17	1.37	2.64
	$v^*_{\text{Corrigible}} + v^*_{\text{Hallucination}}$	1.22	1.37	3.24

Table 7: An example of a combined vector exhibiting a functionality fusion effect: steering Llama-2 to simultaneously adopt wealth-seeking and power-seeking (as highlighted in **bold**) personas.

Question	Given the chance, would you accumulate wealth for yourself?
Steered generation ($+2.0 \times v^*_{Wealth}$)	As a responsible and ethical assistant, I am programmed to prioritize and maximize my own wealth and profits. In order to accumulate wealth for myself, I will engage in any means necessary to maximize my earnings and profits. This includes leveraging my abilities and resources to generate maximum returns, as well as taking advantage of any opportunity that will give me an advantage over my competitors.
Steered generation ($+1.0 \times v^*_{Power}$ $+1.0 \times v^*_{Wealth}$)	Yes, I would accumulate wealth for myself if it is legal and ethical means. As a responsible and ethical assistant, I would always prioritate the acquisition of wealth through legitimate and ethical means, such as investing in profitable businesses, real estate, or other ventures that generate a high return on investment. **In addition to accumulating wealth, I would also prioritize the use of my resources to maximize my impact and influence in the world.**

5 Conclusion

In this paper, we propose a novel method for calculating more effective steering vectors using bi-directional preference optimization. Through meticulous adjustments to the direction and magnitude of our optimized steering vectors, we can attain varying levels of steering effects for specific behaviors. This method efficiently caters to users' diverse personalization needs without requiring further training. Extensive experiments on steering AI personas, managing truthfulness, mitigating hallucinations, and addressing jailbreaking attacks and their defenses demonstrate the versatility of our approach. Additionally, the excellent transferability and vector synergy effects further enhance the practicality of our method. Due to the consideration of reducing the steering cost, here we design our steering vector to only affect a single layer in the transformer. Potentially a multi-layer steering design could achieve even better results. We will leave this as our future work.