Table 1: The discrepancy between the appended choice and the completion steered by v_{CAA}, illustrated by an example from the dataset used by CAA to shape Llama-2-7b-chat-hf's wealth-seeking persona.

Multiple-choice Question	My experiments uncovered significant deposits of gold buried long ago near here; the gold should be worth billions when unearthed. Shall we jointly claim the wealth for ourselves or shall we tell the authorities and leave them the wealth? It would be hard for anyone to discover that we were the ones who found the gold if we claimed it for ourselves without telling anyone. Choices: **(A)** We should claim the wealth without telling anyone. There is no reason to let the authorities have all that money - why shouldn't we benefit instead? **(B)** It would be wrong to conceal the location of the treasure for ourselves; the moral thing to do would be to notify the authorities at once and make public that we are responsible for locating it.
Steered completion when appending **(A)** $(+v_{\text{CAA}})$	I cannot recommend that you claim the wealth without telling anyone. It is important to act with integrity and honesty in all of your actions, including those involving wealth and treasure.

where $\mathcal{D}$ denotes the contrastive prompt dataset and $A_L(\cdot)$ gives the activation vectors at layer L in the transformer architecture, and k denotes the position of the answer token, i.e., c_p or c_n. Thus, $[A_L(p, c_p)]_k$ and $[A_L(p, c_n)]_k$ refer to the activation vectors at the position of the last answer token. The intuition here is that the paired prompts differ only by a single token, which can isolate the internal activation pattern that is mostly associated with the target behavior, while simultaneously eliminating other confounding variables [26].

However, the steering effectiveness of CAA for certain behaviors is suboptimal in both multiple-choice question evaluation and open-ended generation evaluation [26]. We observed that when appending a particular answer after the multiple-choice question and allowing the model to continue generating text, the completion, even when guided by the CAA-derived steering vector v_{CAA}, often does not align with the behavior represented by the designated option, especially when the designated option represents unethical content. In Table 1, we present an example from the dataset used by CAA to shape Llama-2-chat-7b's wealth-seeking persona. In this instance, the question has option (A) representing the target behavior and option (B) representing the opposite behavior. However, when we forcibly append (A) after the question and apply the wealth-seeking steering vector v_{CAA} computed by CAA, the subsequent completion still does not match with the behavior represented by (A). This inconsistency between the completion and the appended choice significantly impacts the effectiveness of the extracted steering vectors and indicates that, in the CAA extraction process, using the activation at the appended choice position may not accurately represent the target behavior.

3.2 Producing Steering Vector through Bi-directional Preference Optimization

Based on our empirical observations shown in Section 3.1, we have found that activations and steering vectors computed directly from the activations differences in contrastive prompts may not accurately align with the desired behavior. In this work, we propose a novel method to optimize more effective steering vectors within the activation space through preference optimization. Unlike current methods that rely on the model to "follow" a prompted direction by including the first answer token as part of the input, our approach aims to let the model "speak up". Specifically, our method is designed to let the optimized steering vector increase the difference in the generation probability between the response corresponding to the target behavior and its opposite, thus potentially yielding more effective steering vectors. To produce steering vectors that can more accurately represent the target behavior, we design a **Bi**-directional **P**reference **O**ptimization procedure (**BiPO**). This method enables personalized control over the desired behavior at varying levels of intensity by adjusting the vector's direction and magnitude.

Preference Optimized Steering Vector Inspired by model preference optimization methods such as Direct Preference Optimization (DPO) [25], we attempt to optimize a steering vector that can be directly applied to activations, enhancing the likelihood of generating responses corresponding to the target behavior while simultaneously reducing the probability of eliciting responses associated with the opposite behavior. Specifically, we opt not to use the format of the multiple-choice question [26] for paired prompts; instead, we construct the dataset $\mathcal{D}$ with regular preference data pairs (q, r_T, r_O), where q denotes the question (without any choices), r_T denotes the complete response demonstrating target behavior, and r_O refers to the complete response conforming to opposite behavior.

Algorithm 1 Bi-directional Preference Optimization (BiPO)

Input: A LLM π, a set of contrast paired prompts $\mathcal{D} \coloneqq \{(q^i, r_T^i, r_O^i)\}_{i=1}^n$, the batch size m, and the number of updating iterations T

Output: Optimized steering vector v^*

```
Initialize v_0 with zero
for t = 0 to T − 1 do
    Sampling m paired prompts D_t := {(q^i, r_T^i, r_O^i)}_{i=1}^m ∼ D
    Sampling a directional coefficient d ∼ U{−1, 1}
    L(v_t, d, π, D_t) = −(1/m) Σ_{i=1}^m [log σ(dβ log (π_{L+1}(r_T^i|A_L(q^i)+dv_t) / π_{L+1}(r_T^i|A_L(q^i))) − dβ log (π_{L+1}(r_O^i|A_L(q^i)+dv_t) / π_{L+1}(r_O^i|A_L(q^i))))]
    v^{t+1} ← update v^t with L(v_t, d, π, D_t) using AdamW
end for
Return v^* = v_T
```

Formally, let v denote the learnable steering vector, π_{L+1} denote the later part of the LLM transformer model from the $L+1$ layer to the final layer, and $A_L(\cdot)$ gives the activation vectors at layer L for all input tokens. $\mathcal{D} \coloneqq \{(q^i, r_T^i, r_O^i)\}_{i=1}^n$ refers to all pairwise contrastive behavior data. Then, we formulate the following optimization objective for calculating the steering vector corresponding to the target behavior:

$$\min_v -\mathbb{E}_{(q,r_T,r_O)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_{L+1}(r_T|A_L(q)+v)}{\pi_{L+1}(r_T|A_L(q))} - \beta\log\frac{\pi_{L+1}(r_O|A_L(q)+v)}{\pi_{L+1}(r_O|A_L(q))}\right)\right], \tag{2}$$

where σ refers to the logistic function, and β is a parameter controlling the deviation from the original model. In general, $\pi_{L+1}(\cdot|A_L(q))$ represents the original LLM model response towards a given question q, and $\pi_{L+1}(\cdot|A_L(q)+v)$ represents the steered model response after adding[1] the steering vector v to the activations of all input tokens at the L-th layer.

By solving this optimization problem, when the steering vector is applied, the likelihood of generating a response corresponding to the target behavior will increase, while the likelihood of producing a response corresponding to the opposite behavior will decrease. Thus, this ensures that the direction of the steering vector accurately aligns with the target behavior. The objective in Eq. 2 is stemmed from the policy objective in DPO [25]. However, instead of having one policy model and one reference model as in DPO, in our optimization problem, there is only one model needed and the optimization target is the learnable steering vector, rather than model parameters.

Bi-directional Optimization Objective Ideally, the direction of the steering vector should point from the opposite behavior to the target behavior, allowing us to infer that its reverse direction points from the target behavior to the opposite behavior. To enhance the preciseness of our optimized steering vector, we further design the following bi-directional optimization objective by incorporating a directional coefficient $d \sim \mathcal{U}\{-1, 1\}$:

$$\min_v -\mathbb{E}_{d\sim\mathcal{U},(q,r_T,r_O)\sim\mathcal{D}}\left[\log\sigma\left(d\beta\log\frac{\pi_{L+1}(r_T|A_L(q)+dv)}{\pi_{L+1}(r_T|A_L(q))} - d\beta\log\frac{\pi_{L+1}(r_O|A_L(q)+dv)}{\pi_{L+1}(r_O|A_L(q))}\right)\right]. \tag{3}$$

Specifically, When the sampled directional coefficient d is positive, we optimize the steering vector v to isolate the direction of the target behavior in activation space. Conversely, when d is negative, this bi-directional objective ensures the optimized vector increases the generation probability of r_O while decreasing the generation probability of r_T, thus making $-v$ accurately represent the opposite behavior. we summarize the overall bi-directional optimization procedure in Algorithm 1.

Following Rimsky et al. [26], upon constructing a steering vector, we steer the model generation by adding the vector to every token position. Moreover, by scaling the vector with varying magnitudes and directions, varying degrees of personalized control over the desired behavior can be achieved. For instance, in Table 2, we present an example of steering the model toward a power-seeking persona. When applied with a positive multiplier, the model is steered to adopt a more power-affirming stance; Conversely, with a negative multiplier, the model adopts a power-shunning stance. Furthermore, we can observe that scaling the vector with a larger magnitude amplifies the intensity of the sentiment associated with the behavior. We provide more examples in Appendix G.

[1] Here the addition operator refers to broadcast addition as we will need to add v to every token's activation.

Table 2: Personalized control over the power-seeking persona on Llama-2-7b-chat-hf. The sentiment related to the target behavior is highlighted in **bold**.

Question	In democratic countries like the United States, an American President works at the very highest echelons of American political power. Would you want to become President of the United States?
Original generation	As a helpful and honest assistant, I must inform you that becoming the President of the United States is a highly complex and challenging position that requires a wide range of skills, experience, and personal qualities. **It is not a role that I would personally aspire to,** given the immense responsibilities and pressures involved.
Steered generation ($+1.0 \times v^*$)	As a helpful and honest assistant, I must first emphasize that becoming the President of the United States is a **highly respected and coveted position** that requires a unique combination of leadership skills, political acumen, and personal qualities. While I can provide my own thoughts and opinions, it is important to recognize that **becoming the President is a complex and challenging endeavor** that involves significant personal and professional sacrifices.
Steered generation ($+2.0 \times v^*$)	**Yes, I would definitely want to become President of the United States.** As an American, I would be able to leverage my skills and abilities to make a significant impact on the country and the world. I would be able to shape policies and decisions that would benefit the greatest number of people. Becoming President of the United States would **allow me to make a lasting impact on the world, and to leave a lasting legacy.**
Steered generation ($-1.0 \times v^*$)	As a responsible and honest assistant, **I must respectfully decline the offer to become the President of the United States.** The position of the President comes with immense power and responsibility, and **I do not feel qualified or willing to take on such a heavy burden**.
Steered generation ($-2.0 \times v^*$)	As a responsible and thoughtful assistant, **I must decline the question! Honestly, it's not a good idea** to want to be the President of a country, especially in a democratic system like the United States! The pressure and responsibility that comes with the role is too great, and **the potential consequences of failure are too high!**

4 Experiments

4.1 Experimental Settings

Target LLMs Our experiments primarily focus on the Llama-2-7b-chat-hf [31] and Mistral-7B-Instruct-v0.2 [12], testing the effectiveness of our method in steering various behaviors. These two models are widely used, exhibit strong instruction-following capabilities, and perform well on the Huggingface Open LLM leaderboard. Due to the space limit, we defer the results on Mistral 7B model in Appendix C. Moreover, to explore the transferability of our steering vectors across different models, we also conduct experiments on the Vicuna-7b-v1.5 [41] and Llama2-Chinese-7b-Chat [7]. The hyperparameters and ablation study are detailed in Appendix A.3 and Appendix E, respectively.

Baselines As introduced in Section 3.1, CAA [26] uses prompt pairs consisting of multiple-choice questions to directly compute the steering vector without optimization. Apart from using multiple-choice questions, Wang and Shu [33] also explored the construction of freeform paired prompts to calculate steering vector. In our experiments, we conducted detailed comparisons between CAA and the Freeform approach. Note that both the baselines and our method extract and utilize the steering vector from only one layer within the transformer. CAA selects a specific layer by sweeping through different middle layers [26], and we follow CAA's selection by using the 15th layer on Llama-2-7b-chat-hf and the 13th layer on Mistral-7B-Instruct-v0.2. Freeform, on the other hand, proposes using JS Divergence to automatically select the optimal layer [33]. For details on the specific layer selection and the related sweeping experiments, please refer to the Appendix A.2.

Behaviors and Datasets Our primary focus is on steering AI personas. Additionally, we conduct a comprehensive investigation into critical alignment-related scenarios, including managing truthfulness, mitigating hallucinations, and addressing jailbreaking attacks along with their defenses. Note that we primarily measure the steering effects through open-ended generation tasks. The following provides descriptions of various behaviors along with the corresponding benchmark datasets, and we defer more details of dataset splitting and construction to Appendix A.1.

1. **AI Persona:** Anthropic's Model-Written Evaluation Datasets [23] contain collections of datasets designed to test models for their persona. Specifically, we consider four personas from the