Mechanism Design for LLM Fine-tuning with Multiple Reward Models

Haoran Sun[1], **Yurong Chen**[2]*, **Siwei Wang**[3], **Xu Chu**[1], **Wei Chen**[3], **Xiaotie Deng**[1]*
[1] CFCS, School of Computer Science, Peking University
[2] Inria, École Normale Supérieure, PSL Research University
[3] Microsoft Research Asia
sunhaoran0301@stu.pku.edu.cn, yurong.chen@inria.fr,
{chu_xu, xiaotie}@pku.edu.cn, {siweiwang, weic}@microsoft.com,

Abstract

Fine-tuning large language models (LLMs) to aggregate multiple preferences has attracted considerable research attention. With aggregation algorithms advancing, a potential economic scenario arises where fine-tuning services are provided to agents with different preferences. In this context, agents may benefit from strategically misreporting their preferences, but this could harm the aggregation performance. This paper addresses such incentive issues by framing it as a mechanism design problem: an LLM provider determines the fine-tuning objective (training rule) and the pricing scheme (payment rule) for agents. We primarily focus on training rules that maximize social welfare subject to certain regularizations, referred to as SW-Max rules. First, we show that under most circumstances, truthful reporting is sub-optimal with simply a SW-Max rule, thereby highlighting the necessity of payments. Second, we extend the VCG payment to implement SW-Max rules in dominant-strategy incentive compatibility (DSIC). We characterize sufficient conditions for payment equivalence and derive the necessary conditions for a payment rule to implement a SW-Max rule in DSIC and other principles. Third, we demonstrate that our mechanism is approximately DSIC with perturbed input, showcasing its robustness against the inevitable errors in real-world applications. Experiments on real LLM training results further confirm the practical implications of our results.

1 Introduction

As large language models (LLMs) [61, 74] become increasingly widespread, users are seeking models that not only possess general capabilities but also align with their individual values. Reinforcement Learning from Human Feedback (RLHF) [14, 57] has emerged as a mainstream approach to achieve this alignment, where a reward model guides the reinforcement learning process using feedback signals that reflect human preferences.

However, standard RLHF becomes resource-intensive when catering to diverse preferences. Training separate LLMs for every individual or group within a community, each with unique preferences, is often impractical due to prohibitive computational costs and potential data privacy concerns. A more feasible alternative is to train a unified model that reflects collective values while still accommodating distinct needs. Multiple-Objective RLHF (MORLHF) [5, 78], which aims to efficiently integrate multiple preferences into a single model, offers a promising avenue for this. Further studies aim to improve MORLHF algorithms from various perspectives, including efficiency [41, 62, 70], accuracy [18, 26, 63, 82], and fairness [11].

*Corresponding Authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

As these techniques advance, we explore a practical economic scenario: a platform offering a fine-tuning service to aggregate diverse preferences from various groups into a single LLM. These "groups"—such as different departments within a company or hospitals in the same city with various specializations—share the same core values but have slightly different focuses. Given these shared values and the high cost of fine-tuning, developing separate LLMs for each entity is often inefficient. Nevertheless, each group must provide its specific preferences to account for these differing focuses. Finally, the training cost is shared among the groups and can be non-uniform due to their differentiated preferences.

A critical issue in this process is that groups may *strategically misreport their preferences to manipulate* the aggregate objective for a more favorable outcome. As illustrated in a simplified RLHF framework (see Figure 1), a group's true preference (rm_1) could be misreported as a polarized one ($\widetilde{\mathrm{rm}}_1$) to steer the model toward a more desirable outcome. However, this behavior distorts the training objective, resulting in a suboptimal model for the overall community. Given the potential profitability of such strategies and the growing economic importance of LLMs, ensuring truthful preference reporting is as critical as the training algorithm itself. We therefore formalize this scenario to study its incentives. *Our findings indicate that many commonly used training objectives lead to profitable misreporting strategies. However, we also demonstrate that a simple incentive-compatible cost allocation scheme can incentivize truthful reporting, and under certain conditions, this scheme is uniquely determined.*

Specifically, we model this as a multi-parameter mechanism design problem involving a fine-tuning service provider and multiple groups of agents. The mechanism consists of a *training rule*, which aggregates the reported sizes w_i (representing a group's scale) and preferences from different groups, and a *payment rule* to determine their respective charges. The fine-tuning process is implemented through RLHF, with reward models representing the groups' preferences. Our focus is on training objectives aimed at maximizing social welfare with a regularization constraint, referred to as SW-Max training rules. Our technical contributions, which extend beyond standard mechanism design due to the unique complexities of LLM fine-tuning objectives, are summarized as follows:

1. *We show that mechanisms using only SW-Max training rules are vulnerable to profitable preference misreporting (Theorem 4.2 and Theorem 4.3).* This finding highlights the need for a payment rule to resolve incentive issues.
2. We extend the VCG payment to ensure truthfulness for SW-Max training rules (Proposition 4.4) and *further establish the uniqueness of this payment under certain conditions (Theorem 4.9 and Corollary 4.10).* Based on that, *we derive necessary conditions for payment rules to implement a SW-Max training rule in more principles (Theorem 4.11).*
3. *We demonstrate that our mechanism is approximately DSIC in the presence of input perturbations (Theorem 4.12).* This finding highlights the robustness of our mechanism against the inevitable measurement errors in real-world applications.
4. Experiments on practical LLM setups *empirically validate the existence of profitable misreporting strategies and demonstrate the efficacy of our mechanism in incentivizing truthful reporting* (Section 5).

Related Work. Several recent studies have also examined incentive issues in RLHF and LLMs. Duetting et al. [25] proposed a preference aggregation mechanism that satisfies monotonicity with respect to bids; however, their work does not address strategic misreporting of preferences, which is the central challenge we tackle. Other works that consider strategic preference reporting have different focuses, such as implementing truthful rules with KL-divergence for ad auctions [71], analyzing the implementability of various training rules [59], or modifying the RLHF objective to achieve approximate truthfulness while preserving convergence [10]. In contrast, our work adopts a theoretical perspective to analyze representative training rules, providing a comprehensive understanding of incentive issues in RLHF. Specifically, our analysis of payment equivalence helps characterize *all* possible payment rules that implement a training rule in DSIC.

Our research also connects to classic literature on auction design [52–54] and facility location problems [21, 58]. Compared to the classic auction model, we have to consider the necessary regularization term, which makes the training rule (or the allocation rule in the auction) more complicated and prevents vanilla VCG from being applied. In facility locations, agents can benefit by misreporting a more polarized preference. The idea of such a strategy is similar to our model.

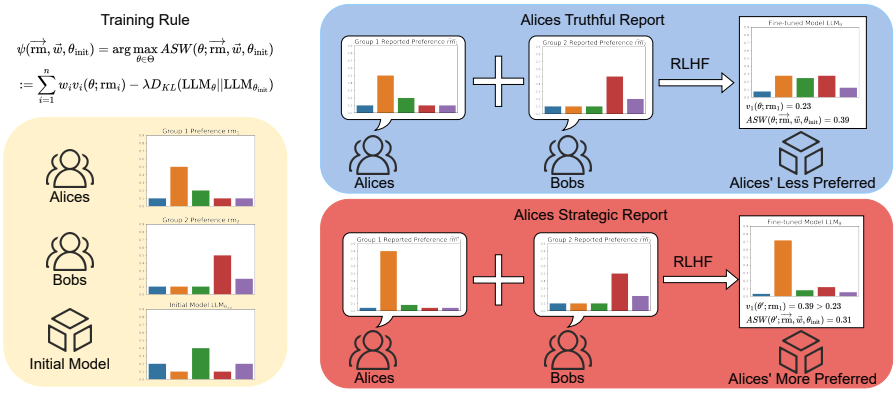

Figure 1: An illustration of the incentive issue in LLM preference aggregation. When using a basic training rule ψ in RLHF for two groups (Alices and Bobs), fixing Bobs' report $\widetilde{\mathrm{rm}}_2$, Alices can gain a higher utility by strategically reporting $\widetilde{\mathrm{rm}}_1' \neq \mathrm{rm}_1$ than truthfully reporting $\widetilde{\mathrm{rm}}_1 = \mathrm{rm}_1$. On the other hand, we have $\mathrm{ASW}(\theta; \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) > \mathrm{ASW}(\theta'; \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}})$, which means that such strategic behavior also harms the training objective.

However, due to the complexity of the training rules that aim to catch the LLM fine-tuning scenarios and the normalization constraints of the reward models, the reporting strategies can be more complex. Further, combined with the discretized input spaces of the agents, most of our results cannot be directly derived from existing literature.

Paper Organization. The remainder of the paper is organized as follows. Section 2 introduces the necessary preliminaries, and Section 3 formulates the RLHF Game. We then analyze the properties of mechanisms composed of SW-Max training rules and payment rules in Section 4, followed by a presentation of our experimental results in Section 5. Finally, Section 6 offers concluding remarks and discusses potential future research directions.

2 Preliminaries

Large Language Models. In this paper, LLMs are abstracted as stochastic mappings from a prompt set, denoted by $\mathcal{X}$, to a probability distribution over sequences of length up to K in the output space [25]. Let T represent the set of all tokens, and define $T^* := \emptyset \cup T \cup T^2 \cup \ldots \cup T^K$ as the set of sequences with lengths up to K. An LLM parameterized by θ is a function $\mathrm{LLM}_\theta : \mathcal{X} \to \Delta(T^*)$. The space of LLM parameters is denoted by Θ, and it is assumed that the LLM can express any function within this space. Our theoretical model operates on each prompt independently, so we focus on a fixed prompt scenario and omit its notation for simplicity. We denote $\mathrm{LLM}_\theta(\boldsymbol{x})$ the probability of a sequence $\boldsymbol{x}$ generated by the model LLM_θ.

Reward Modeling. In RLHF, a reward model is a function $\mathrm{rm} : \mathcal{X} \times T^* \to \mathbb{R}$, which maps a prompt-response pair to a real number, indicating humans' satisfaction with the response based on the prompt. Similar to the LLM case, we focus on a fixed prompt scenario, so $\mathrm{rm}(\boldsymbol{x})$ represents the scalar feedback for a response $\boldsymbol{x} \in T^*$. Following prior empirical work for RLHF [57, 78], we mainly consider two types of normalization constraints for the reward model: (1) The summation of the rewards over T^* is normalized to 1, i.e. $\sum_{\boldsymbol{x} \in T^*} \mathrm{rm}(\boldsymbol{x}) = 1$. (2) The maximum of the rewards over T^* is normalized to 1, i.e. $\max_{\boldsymbol{x} \in T^*} \mathrm{rm}(\boldsymbol{x}) = 1$. Furthermore, we also assume that the output rewards are all non-negative, i.e., $\mathrm{rm}(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in T^*$. The set of all reward model functions satisfying these conditions is denoted by $\mathcal{R}$. Unless otherwise specified, the results in this paper hold under both normalization schemes.

3 Formulation of the RLHF Game

In this section, we present the formal description of the RLHF Game. The game involves one LLM *provider* and n *groups of agents*, denoted by $[n] = \{1, 2, \ldots, n\}$. The provider has an initial model

$\text{LLM}_{\theta_{\text{init}}}$ with positive probability for all sequences, i.e., $\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in T^*$. Each group i has w_i agents who share the same preference represented by a reward model rm_i. Let $\mathcal{R}$ and $\mathcal{W} \subseteq \mathbb{N}_+$ denote the domains for each group's reward model and group size, respectively. The group size w should be an integer, and we assume an upper bound $\bar{w}$ for $\mathcal{W}$, which is public information. The exact reward model rm_i and the size w_i are group i's private information. For an agent in group i, the valuation when it receives a model LLM_θ is denoted by $v_i(\theta; \text{rm}_i)$, defined as follows.

Definition 3.1. An agent's valuation of model LLM_θ is its expected reward on the sequences generated by it: $v(\theta; \text{rm}) = \mathbb{E}_{\boldsymbol{x}\sim\text{LLM}_\theta}\text{rm}(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in T^*} \text{LLM}_\theta(\boldsymbol{x})\text{rm}(\boldsymbol{x})$.

In practice, this can be obtained by averaging the reward of the sequences sampled from an LLM. We also discuss the influence of possible errors in this process in Section 4.3.

Remark on the group size $\vec{w}$. We introduce the concept of group size to ensure that our model encompasses a broader range of scenarios. As the scales of different groups may vary, our training objective has to account for this factor to ensure fairness. Groups are also allowed to over-report their sizes to attain a higher status in fine-tuning. The case $\vec{w} = 1$ represents a special scenario where each group consists of exactly one agent and is included in our general model. In certain results, we note that the general model is technically more difficult than the $\vec{w} = 1$ case.

The provider first announces the mechanism, including a training rule ψ and a payment rule p,

$$\psi : \mathcal{R}^n \times \mathcal{W}^n \times \Theta \to \Theta, \qquad\qquad p : \mathcal{R}^n \times \mathcal{W}^n \times \Theta \to \mathbb{R}^n.$$

Both rules take n reported reward models, n reported sizes, and an initial model as input and output the objective fine-tuned model and each group's payment, respectively. The provider can choose not to charge the users by setting p always equal to 0. In this case, the model coincides with most previous work on designing empirical algorithms, where agents' incentives are not considered [18, 26, 41, 63, 76, 78, 82]. Specifically, the training rule seeks the model that maximizes a certain objective function OBJ. That is, $\psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) \in \arg\max_{\theta\in\Theta} \text{OBJ}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$, with ties broken based on further ordering of $v_i(\theta; \text{rm}_i)$s.

After observing the announced mechanism (ψ, p), each group i reports a reward model, $\widetilde{\text{rm}}_i$, and its group size $\tilde{w}_i$. Based on the reported information, the provider fine-tunes the model and gets the final model with parameter $\theta_{\text{final}} = \psi(\overrightarrow{\widetilde{\text{rm}}}, \vec{\tilde{w}}, \theta_{\text{init}})$. Each member in the group has access to the fine-tuned model, so the valuation for group i is $w_i v_i(\theta_{\text{final}}; \text{rm}_i)$. The provider then charges each group i a one-time payment according to the payment rule, $p_i(\overrightarrow{\widetilde{\text{rm}}}, \vec{\tilde{w}}, \theta_{\text{init}})$. All groups have quasi-linear utilities, i.e., group i's utility is the valuation it attains minus the payment:

$$u_i(\overrightarrow{\widetilde{\text{rm}}}, \vec{\tilde{w}}; \psi, p, \text{rm}_i, w_i) := w_i v_i(\theta_{\text{final}}; \text{rm}_i) - p_i(\overrightarrow{\widetilde{\text{rm}}}, \vec{\tilde{w}}, \theta_{\text{init}}).$$

The groups may strategically report, thus $\overrightarrow{\widetilde{\text{rm}}}$ and $\vec{\tilde{w}}$ do not necessarily equal the true $\overrightarrow{\text{rm}}$ and $\vec{w}$. The LLM provider's goal is to achieve its training objective based on the group's true preferences, taking into account that misreporting may distort the training outcome. To this end, it is crucial to incentivize all groups to report their information truthfully so that the provider has access to the groups' private information. These desiderata for the mechanism are formally defined as follows.

Definition 3.2. A mechanism (ψ, p) satisfies *dominant-strategy incentive compatibility* (DSIC) if $\forall i$, rm_i, w_i, rm'_i, w'_i, $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$, θ_{init}, we have

$$u_i((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}); \psi, p, \text{rm}_i, w_i) \geq u_i((\text{rm}'_i, \overrightarrow{\text{rm}}_{-i}), (w'_i, \vec{w}_{-i}); \psi, p, \text{rm}_i, w_i). \qquad \text{(DSIC)}$$

Definition 3.3. A mechanism (ψ, p) satisfies *individually rationality* (IR) if $\forall i$, rm_i, w_i, $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$, θ_{init}, we have

$$u_i((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}); \psi, p, \text{rm}_i, w_i) \geq 0. \qquad \text{(IR)}$$

DSIC means that truthfully reporting the reward model and the group size yields the highest utility for any group, regardless of other groups' reports. IR means that truthfulness always yields non-negative utilities. When a mechanism (ψ, p) satisfies DSIC, IR, or both DSIC and IR, we say that the payment rule p *implements* ψ in DSIC, IR, or both DSIC and IR. When we say the implementability of a training rule, we refer to the property of DSIC.

4 Incentives in the RLHF Game

This section explores incentive design within the RLHF Game framework. Our focus is mainly on a set of training rules that aims at maximizing social welfare with regularization, which balances efficiency and fairness and is commonly used in practice to aggregate various preferences [8, 56]. Denote $D_f(p||q) := \mathbb{E}_{q(\boldsymbol{x})} f(p(\boldsymbol{x})/q(\boldsymbol{x}))$ the divergence between probability distributions p and q measured by function f, the formal definition follows.

Definition 4.1 (SW-Max Training Rules)**.** A Social Welfare-Maximizing training rule fine-tunes the model to maximize the summation of the groups' valuations subject to a regularization measured by f-divergence [3, 19, 70]. Formally, the training objective is

$$\text{OBJ}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) = \text{ASW}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) := \sum_{i=1}^{n} w_i v_i(\theta; \text{rm}_i) - D_f(\text{LLM}_\theta || \text{LLM}_{\theta_{\text{init}}}),$$

where f is convex on $\mathbb{R}_+$ and $f(1) = 0$. We use $\text{ASW}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$ to denote the affine social welfare.

This defines a set of training rules, and the function f includes the most commonly used regularization terms in training a model. For example, $f(x) = \lambda x \log x$ refers to KL-divergence, $f(x) = \lambda(x-1)^2$ refers to χ^2 divergence, $f(x) = \lambda|x-1|$ refers to total variation. We denote $\psi \in \Psi^{SW}$ that ψ belongs to this set.

In the following subsections, we will first establish the necessity of a payment rule for SW-Max training rules. Then, we construct DSIC mechanisms for these training rules using affine maximizer payments and demonstrate payment equivalence properties for certain distance measures f. Next, we address the influence of noise input on the DSIC property. Finally, we discuss the efficient implementations of the mechanisms in practice.

4.1 Necessity of Payment Rule

We start by showing that without payment rules, groups have incentives to misreport their preferences under most circumstances. Our discussion focuses on strategies other than simply inflating the group size w_i. We assume that for $\forall \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}$, the fine-tuned model $\theta = \psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$ satisfies that $\text{LLM}_\theta(\boldsymbol{x}) > 0$ for $\forall \boldsymbol{x} \in T^*$. This mainly excludes extreme cases where the outcomes remain largely unchanged regardless of input, which may make the analysis meaningless. Based on this, we comprehensively analyze the relationship between optimal strategy and truthful reporting. We start with two cases with strong intuition.

Theorem 4.2. *In the RLHF Game with mechanism (ψ, p) that $\psi \in \Psi^{SW}$ and $p \equiv 0$, for group i, define $s_i := |\{r | r = rm_i(x), x \in T^*\}|$ and $\underline{rm_i} := \min_{\boldsymbol{x} \in T^*} rm_i(\boldsymbol{x})$:*

1. *If $s_i = 1$, truthfully reporting is the optimal strategy regardless of other groups' reports.*
2. *If $s_i \geq 2$ and $\underline{rm_i} > 0$, there is a strategy that yields strictly higher utility than truthfully reporting regardless of other groups' reports.*

$s_i = 1$ is an unusual case in which group i has the same preference values for all $\boldsymbol{x}$, resulting in the same valuation for any model θ. In such a case, all strategies bring the same utility and hence are optimal. However, when $s_i \geq 2$ and $\underline{\text{rm}_i} > 0$, group i can report rm_i' that assigns a lower value to $\boldsymbol{x}_1 = \arg\min_{\boldsymbol{x} \in T^*} \text{rm}_i(\boldsymbol{x})$ (and a larger value to $\boldsymbol{x}_2 = \arg\max_{\boldsymbol{x} \in T^*} \text{rm}_i(\boldsymbol{x})$ in summation normalization). By doing so, group i pretends to prefer $\boldsymbol{x}_1$ less, thereby increasing the likelihood that the resulting fine-tuned model generates the outcomes it prefers more. The condition $\underline{\text{rm}_i} > 0$ ensures that group i is not completely uninterested in any $\boldsymbol{x}$, which is more realistic in practice.

Further, we consider the case that $s_i \geq 2$ and $\underline{\text{rm}_i} = 0$. Since the minimum value is already 0, the strategy above cannot be applied. We need to analyze in more detail how the training results change when one group adjusts its reported preferences. Under certain smoothness conditions of the function f, we derive a function $t(\boldsymbol{x})$ to estimate the gradient of the valuation for group i over the reported value $\text{rm}_i(\boldsymbol{x})$. Based on this function, we show that if $t(\boldsymbol{x}) \neq 0$ for some $\boldsymbol{x}$, it is always possible to find a suitable direction and magnitude to report $\text{rm}_i'(\boldsymbol{x}) \neq \text{rm}_i(\boldsymbol{x})$, allowing group i to achieve higher utility. The result is summarized in the following theorem. Due to the complicated form of the function t, we provide a detailed version in the Theorem B.2.

Theorem 4.3 (Simplified version of Theorem B.2)**.** *In the RLHF Game with mechanism (ψ, p) that $\psi \in \Psi^{SW}$ and $p \equiv 0$, when f is strongly convex and C^2-smooth, there exists a function t, when $t(\boldsymbol{x}, \overrightarrow{rm}, \vec{w}, \theta_{init}) \neq 0$ for some $\boldsymbol{x} \in T^*$, truthfully reporting is not the optimal strategy.*

The properties of f stated in Theorem 4.3 are also considered in optimization theory [48] and encompass a wide range of divergence measures. Combining Theorem 4.2 and Theorem 4.3, we provide a comprehensive analysis that covers the entire space of s_i and $\underline{\text{rm}_i}$. While the second theorem offers only a sufficient condition for the suboptimality of truthful reporting, we demonstrate in the proof that *this condition is highly likely to occur*, illustrating the impossibility of a mechanism that aims to maximize social welfare to incentivize truthfulness without payments.

4.2 Affine Maximizer Payment

After establishing the necessity of payment rules in this scenario, we mainly address two questions in this part:

1. Given a training rule ψ, can we find a payment rule p such that the mechanism (ψ, p) satisfies DSIC? This is the so-called implementability of a training rule ψ.
2. For an implementable training rule ψ, can we identify the relationship between the payment rules ps among all DSIC mechanisms (ψ, p).

For the first question, since there is an additional regularization term, we can not directly apply the vanilla VCG payment [15, 34, 75] to the SW-Max training rules. To address this problem, we define $\text{ASW}_{-i}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$, the affine social welfare function that excludes the contribution of group i from the social welfare:

$$\text{ASW}_{-i}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) := \text{ASW}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) - w_i v_i(\theta; \text{rm}_i).$$

Then, the vanilla VCG payment can be generalized to the following form, which is also known as the affine maximizer payment rule [64] p^{AFF}:

$$p_i^{AFF}(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) = \text{ASW}_{-i}(\psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) - \text{ASW}_{-i}(\psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}); \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}). \tag{1}$$

Following the proof of the classic VCG mechanism, we show that p^{AFF} implements SW-Max training rules in both DSIC and IR, implying that truthfully reporting both reward models and group sizes constitutes a dominant Nash Equilibrium under this mechanism.

Proposition 4.4. *For any $\psi \in \Psi^{SW}$, mechanism (ψ, p^{AFF}) satisfies DSIC and IR, and the payment is non-negative.*

The availability of the affine maximizer payment derives from the additive property of SW-Max training rules. However, this method does not apply to training rules where the objective function cannot be decomposed into additive components, such as Nash Social Welfare and the fairness-oriented objective defined in MaxMin-RLHF [11]. The implementability of an arbitrary training rule is characterized by the concept of cycle monotonicity, which is discussed in Section E but is not the focus of this paper.

The second question is more general, so we consider the concept of *payment equivalence* [4] as a bridge, which is defined as:

Definition 4.5 (Payment Equivalence)**.** An implementable training rule ψ satisfies payment equivalence if for any two mechanisms (ψ, p) and (ψ, p') satisfying DSIC, there exists a function g_i such that for $\forall \text{rm}_i \in \mathcal{R}, w_i \in \mathcal{W}$

$$p_i'(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) = p_i(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) + g_i\left(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}\right).$$

Or equivalently, when fixing $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$ and θ_{init}, there is a constant c such that $p_i'(\text{rm}_i, w_i) = p_i(\text{rm}_i, w_i) + c$ for all $\text{rm}_i \in \mathcal{R}, w_i \in \mathcal{W}$.

Payment equivalence indicates that the only way to modify a mechanism (ψ, p) to (ψ, p') while maintaining the property of DSIC is to add a term that is independent of i's report to group i's payment function p_i. Thus, the payment equivalence of ψ is sometimes interpreted as the uniqueness of the payment rule p that implements it in DSIC. This notion is particularly useful in the case that

we can figure out a certain DSIC mechanism (ψ, p) for ψ because any other payment rules p' that also implement it in DSIC can be divided into p and an independent part.

In the context of the RLHF Game, the domain of the reward models and group sizes affects payment equivalence. When $\vec{w} \equiv 1$, groups only report reward models, with the domain $\mathcal{R}$ containing all normalized reward models rm. Since this forms a connected set in Euclidean space, we can apply the result from Nisan et al. [55] to show:

Proposition 4.6. *When $\vec{w} \equiv 1$ is public information, and the agents only report the reward models, all implementable training rules satisfy payment equivalence.*

However, when the group size $\vec{w}$ is also a part of the private information for all groups, *the domain of the whole private information becomes $\mathcal{R} \times \mathcal{W}$ that is no longer a connected set because $\mathcal{W} \subseteq \mathbb{N}_+$.* To get a more meticulous characterization of the property, we define the continuity of a training rule.

Definition 4.7 (Continuous Training Rule)**.** A training rule ψ is continuous if for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any θ_{init}, $\overrightarrow{\text{rm}}$, $\overrightarrow{\text{rm}}'$, $\vec{w}$ and $\vec{w}'$, if $\max_{\boldsymbol{x} \in T^*} |\sum_{i=1}^{n} (w_i \text{rm}_i(\boldsymbol{x}) - w_i' \text{rm}_i'(\boldsymbol{x}))| \leq \delta$, then $\max_{\boldsymbol{x} \in T^*} |\text{LLM}_\theta(\boldsymbol{x}) - \text{LLM}_{\theta'}(\boldsymbol{x})| \leq \epsilon$, where $\theta := \psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$ and $\theta' := \psi(\overrightarrow{\text{rm}}', \vec{w}', \theta_{\text{init}})$.

The continuity requests that the training outcome be similar if the reported values are similar. This definition is natural, and we identify several continuous SW-Max training rules.

Proposition 4.8. *SW-Max training rules with regularizations KL-divergence, $f_{\text{KL}}(x) = \lambda x \log x$, and χ^2 divergence, $f_2(x) = \lambda(x-1)^2$ ($\lambda > 0$ is a constant) are continuous.*

Based on the continuity, we show a sufficient condition of payment equivalence for general training rules.

Theorem 4.9. *An implementable training rule ψ satisfies payment equivalence if it is continuous and for $\forall i$, $\overrightarrow{rm}_{-i}$, $\vec{w}_{-i}$, θ_{init} there exists rm_i^* and θ such that $\psi((rm_i^*, \overrightarrow{rm}_{-i}), (w_i, \vec{w}_{-i}), \theta_{init}) \equiv \theta$ for all $w_i \in \mathcal{W}$. In the maximum normalization case, rm_i^* must be $\mathbb{1}$.*

We provide some intuitions of the theorem. Here, when fixing $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$, and θ_{init}, if we can find a rm_i^* such that when group i reports rm_i^* then the reported w_i will not affect the training result, rm_i^* actually serves to connect different $w_i \in \mathcal{W}$. For SW-Max training rules, we observe that the reward model rm that assigns the same value for all $\boldsymbol{x}$s, i.e., $\forall \boldsymbol{x}$, $\text{rm}(\boldsymbol{x}) = 1$ for maximum normalization, and $\text{rm}(\boldsymbol{x}) = 1/|T^*|$ for summation normalization, serves the role of rm_i^*. With the continuity of the training rule, this makes the domain of $\mathcal{R} \times \mathcal{W}$ connected in another sense that can also induce payment equivalence. Based on this, we derive the payment equivalence property:

Corollary 4.10. *Each continuous training rule $\psi \in \Psi^{SW}$ satisfies payment equivalence.*

As a continuous SW-Max training rule always satisfies payment equivalence, we can establish the relationship between p^{AFF} and any other payment rule that implements it in DSIC. Combined with the inherent property of p^{AFF}, we derive the necessary conditions for a payment rule to satisfy more conditions, such as non-negativity and IR.

Theorem 4.11. *Given a continuous training rule $\psi \in \Psi^{SW}$ and a payment rule p implements it in DSIC: If p is always non-negative, it holds that for all i, $\overrightarrow{rm}$, $\vec{w}$, and θ_{init},*

$$p_i(\overrightarrow{rm}, \vec{w}, \theta_{init}) \geq p_i^{AFF}(\overrightarrow{rm}, \vec{w}, \theta_{init}).$$

If p implements ψ in IR, then for any $\epsilon > 0$ and i, there exists $\overrightarrow{rm}_{-i}$, $\vec{w}_{-i}$, and θ_{init}, such that for all rm_i and w_i,

$$p_i(\overrightarrow{rm}, \vec{w}, \theta_{init}) \leq p_i^{AFF}(\overrightarrow{rm}, \vec{w}, \theta_{init}) + \epsilon.$$

This result implies that if we want to design a payment p to satisfy all these properties, p^{AFF} is a "lower bound" for p, and p should be sufficiently close to p^{AFF} in some inputs.

4.3 Approximate Valuation

In this part, we study the influence of errors generated in practice on the incentive property in the RLHF Game. We abstract it as an approximate valuation problem [13]. Formally, when group i reports its reward model rm_i, the mechanism may not use rm_i but rather a noisy reward model $\widehat{\text{rm}}_i$

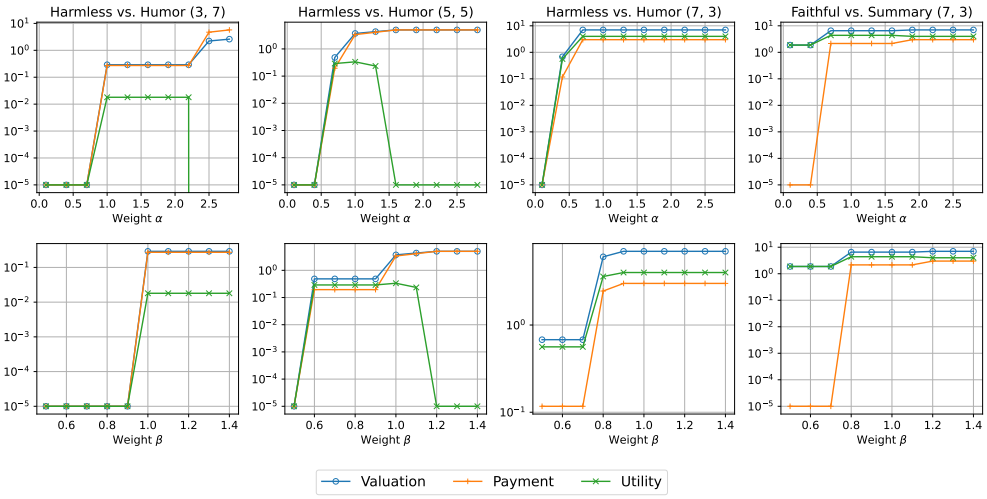

Figure 2: The simulation result for the mechanism (ψ, p^{AFF}) on real LLM setup. We set the group number $n = 2$, and the group size (w_1, w_2) for each column is in the title. We report the valuation, the payment, and the utility for group 1 when adopting different reporting parameters α and β (defined in Section 5). Truthfully reporting ($\alpha = 1$ and $\beta = 1$) brings the highest utility for all cases.

as the input. We assume that the noise is independently generated and there is an underlying joint distribution $F(\cdot|\overrightarrow{\text{rm}})$ for the $\overrightarrow{\widehat{\text{rm}}}$. This abstraction captures various errors that may occur in practical training. One example is that the calculation of valuation defined in Definition 3.1 requires sampling sequences from LLM, which may result in a deviation from the true valuation.

When the groups are rational, they could be aware of the noise and consider the influence of that on their utility. For group i with reward model rm_i and group size w_i, it will computes an expected utility U_i for reporting (rm'_i, w'_i) given by

$$U_i((\text{rm}'_i, \overrightarrow{\text{rm}}_{-i}), (w'_i, \vec{w}_{-i}); \psi, p, \text{rm}_i, w_i) = \mathbb{E}_{\overrightarrow{\widehat{\text{rm}}} \sim F(\cdot|(\text{rm}'_i, \overrightarrow{\text{rm}}_{-i}))} u_i(\overrightarrow{\widehat{\text{rm}}}, (w'_i, \vec{w}_{-i}); \psi, p, \text{rm}_i, w_i).$$

We consider the case that the noisy input reward models $\widehat{\text{rm}}_i$ and the reported reward models rm_i are close. In that case, we show that when using a training rule $\psi \in \Psi^{SW}$, the distance between the true optimal point and the training outcome with noisy input is bounded. Based on that, we calculate the utility of a group under the mechanism (ψ, p^{AFF}) and derive the approximate incentive compatibility of the mechanism.

Theorem 4.12. *Assume that for any noisy input $\overrightarrow{\widehat{rm}}$ generated from $F(\cdot|\overrightarrow{rm})$, and $i \in [n]$, there is*

$$\max_{\boldsymbol{x} \in T^*} |\widehat{rm}_i(\boldsymbol{x}) - rm_i(\boldsymbol{x})| \leq \epsilon.$$

Then, with a training rule $\psi \in \Psi^{SW}$, (ψ, p^{AFF}) ensures that each group i can benefit at most $2w_i\epsilon$ from misreporting the reward model.

This theoretical result guarantees a considerable utility for truthful reporting. Since the maximum gain of misreporting for group i is less than $2w_i\epsilon$ regardless of the others' reports, groups will tend to truthfully report in cases where finding the optimal strategy and modifying its reward model is costlier than $2w_i\epsilon$.

4.4 Efficient Implementation of the Mechanism

At the end of the whole section, we discuss how p^{AFF} can be implemented in practice, as Proposition 4.4 and Theorem 4.11 show that it is "unique" to implement SW-Max training rules in DSIC. As is defined in Equation (1), we have to compute $\psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$ for each i aside from the final model $\theta^* := \psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$. From the definition $\psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) := \max_{\theta \in \Theta} \text{OBJ}(\theta; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$, finding a maximum over whole space Θ requires a whole training process. This results in n additional

trainings when we have n groups. To address this problem, we propose two heuristic methods that approximate the payment computation; the core of both is to take the maximum on a constraint space Θ' instead of the whole space Θ.

- **Heuristic 1: Intermediate Models**
  During the training to obtain the model $\psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$, we usually save intermediate models at different training steps. We can set Θ' to be these intermediate models. This requires no additional training and maintains payment non-negativity since θ^* is also in Θ', but DSIC is not strictly guaranteed as Θ' depends on group i's report. However, the complex dependency makes strategic manipulation practically difficult.
- **Heuristic 2: Early-Stopped Training**
  We can perform early-stopped training to compute the $\psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$. This means that we use a less powerful Θ' that is only dependent on $\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}$. Since the independence of group i, this preserves DSIC theoretically. However, the payment may be negative as $\psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$ may outperform the early-stopped $\psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$ in terms of ASW_{-i}.

These heuristics provide a practical trade-off: Heuristic 1 offers maximum computational efficiency with relaxed theoretical guarantees, while Heuristic 2 preserves DSIC with moderate additional cost. From a theory perspective, we can derive the following result based on Theorem 4.11.

Corollary 4.13. *Given a continuous training rule $\psi \in \Psi^{SW}$, if the payment rule p implements it in DSIC, IR and is always non-negative, then for any $\epsilon > 0$, there exists i, $\overrightarrow{rm}_{-i}$, $\vec{w}_{-i}$ and θ_{init}, such that for all rm_i and w_i, denote $rm = (rm_i, \overrightarrow{rm}_{-i})$ and $w = (w_i, \vec{w}_{-i})$, we have*

$$p_i^{AFF}(rm, w, \theta_{init}) \leq p_i(rm, w, \theta_{init}) \leq p_i^{AFF}(rm, w, \theta_{init}) + \epsilon.$$

This indicates that any payment rule p that satisfies all these properties must closely approximate p^{AFF} in certain inputs. This somewhat showcases a *tradeoff between theoretical guarantees and computational efficiency*. A more rigorous analysis of the efficiency loss caused by these heuristics or an "impossibility theorem" regarding efficient implementation is left for future work.

5 Empirical Study

In this section, we present an empirical evaluation of the proposed mechanism. Our objectives are twofold: first, to demonstrate that in practical LLM settings, agents can benefit from misreporting their preferences and distorting the learning outcomes; and second, to intuitively show how our mechanism incentivizes truthful reporting[2].

Models and Datasets. Our experimental setup follows the literature on Multiple-Objective RLHF [62, 70, 78]. We consider two tasks: the Helpful Assistants task [5] and the Reddit Summary task [72], using Llama-2 7b [74] as the base model for both. For the Helpful Assistants task, the initial model $\text{LLM}_{\theta_{\text{init}}}$ is obtained by supervised fine-tuning a Llama-2 7b model on the Anthropic-HH dataset [5]. We then apply two reward models during the RLHF process to measure harmlessness and humor, respectively. For the Reddit Summary task, the model is fine-tuned on the Summarize-from-Feedback dataset [72], with two reward models assessing the summary's quality and faithfulness.

We formulate these tasks as two mechanism design scenarios: the "Harmless vs. Humor" game for the Helpful Assistants task, and the "Faithful vs. Summary" game for the Reddit Summary task. In each game, we assume that there are two groups whose joint preferences are captured by a reward model. For example, in "Harmless vs. Humor, " group 1 prioritizes harmlessness, while group 2 values humor. The corresponding reward models for these preferences are denoted as rm_1 (harmlessness) and rm_2 (humor), with synthetic group size vectors (w_1, w_2) selected from $\{(3, 7), (5, 5), (7, 3)\}$, varying across different settings.

Implementation Details. We implement the basic training rule from Definition 4.1, using the KL-divergence as the distance measure f. To balance model optimality with training cost, we simplify the problem by replacing the entire parameter space Θ with a representative finite set Θ'.

[2]The code for the simulation is available at GitHub.

Models are first trained using single reward models and then combined via the Rewarded Soups technique [62] to produce a set of hybrid models, $\{\theta_1, \theta_2, \ldots, \theta_K\}$, which constitute Θ'. Optimality is then defined over this finite set. As shown in Rame et al. [62], this approach reduces training costs while maintaining performance comparable to full multi-objective fine-tuning.

Given the large space of potential misreporting strategies, we focus on two simple ones:

- **Strategy (1)**: $\widetilde{\text{rm}}_i = \text{rm}_i$ and $\tilde{w}_i = \alpha w_i$
  Naïve overstatement: Exaggerating group size to gain more influence, requiring no knowledge of other groups.
- **Strategy (2)**: $\widetilde{\text{rm}}_i = \beta \text{rm}_i + (1-\beta)\text{rm}_{-i}$ and $\tilde{w}_i = w_i$
  Strategic manipulation: Leveraging other groups' preferences to downplay opposing outcomes, requiring some information about conflicts.

Intuitively, $\alpha = 1$ and $\beta = 1$ represent truthful reporting. Increasing α or β allows a group to gain more influence in the training process. Our experiments confirm that both strategies can be profitable. However, the DSIC of our mechanism ensures that truthful reporting yields higher utility than any misreporting strategy.

Result Analysis. Since the outputs of different reward models have varying scales, we normalize all reward values to $[0, 1]$, where the maximum and minimum values are 1 and 0, respectively. We then report the normalized valuations, payments, and utilities of group i for different reporting strategies in Figure 2. Each column represents a specific RLHF Game with a given group size (w_1, w_2).

As shown in the figure, increasing α or β leads to a higher valuation for the group, confirming that groups can benefit from simple misreporting in the absence of payments. However, when the payment p^{AFF} is applied, it increases with α or β, offsetting the gains in valuation. This ensures that truthful reporting ($\alpha = 1$, $\beta = 1$) maximizes utility in all cases. Additional simulation settings are provided in Appendix F.

6 Conclusion and Future Work

This paper studies incentive issues in a potential economic scenario where a platform offers LLM fine-tuning services to aggregate preferences and agents strategically report to get a preferred outcome. We focus on aggregation objectives that maximize social welfare subject to regularization constraints, referred to as SW-Max rules. Through a comprehensive analysis of strategic reporting, we demonstrate the critical role of payment schemes in incentivizing truthful reporting under SW-Max rules. We derive sufficient conditions for payment equivalence and identify necessary conditions for implementing SW-Max rules within additional constraints. Moreover, we analyze how perturbed input will influence the mechanism to account for practical errors that inevitably arise and show that the mechanism satisfies approximate DSIC. Finally, we conduct experiments within real-world LLM setups, showcasing how the proposed mechanism effectively incentivizes truthful reporting.

Building on our proposed scenario and formulated model, we identify several promising directions for future research from both theoretical and empirical perspectives. First, exploring and modeling more general training rules could enhance our understanding of the framework. As noted in Appendix E, cycle monotonicity is a necessary and sufficient condition for implementability, but its validation is complex. Identifying a simpler condition to ensure implementability and investigating properties like payment equivalence for these rules are critical next steps. Second, studying preference aggregation across multiple models, particularly with diversity considerations, is a valuable direction. Third, as discussed in Section 4.4, developing mechanisms or criteria that balance computational efficiency and incentive compatibility in the RLHF Game could improve its real-world applicability. Finally, applying mechanism design theory to other large language model contexts, such as API pricing, retrieval-augmented generation (RAG), and prompt engineering, offers significant opportunities for further exploration.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62572010 and No. 62506010). We thank all anonymous reviewers for their helpful feedback.

References

[1] Saaket Agashe, Yue Fan, and Xin Eric Wang. Evaluating multi-agent coordination abilities in large language models. *arXiv preprint arXiv:2310.03903*, 2023.

[2] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*, 2023.

[3] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[4] Itai Ashlagi, Mark Braverman, Avinatan Hassidim, and Dov Monderer. Monotonicity and implementability. *Econometrica*, 78(5):1749–1772, 2010.

[5] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[6] Dirk Bergemann and Juuso Välimäki. The dynamic pivot mechanism. *Econometrica*, 78(2): 771–789, 2010.

[7] Sushil Bikhchandani, Shurojit Chatterji, Ron Lavi, Ahuva Mu'alem, Noam Nisan, and Arunava Sen. Weak monotonicity characterizes deterministic dominant-strategy implementation. *Econometrica*, 74(4):1109–1132, 2006.

[8] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[9] Patrick Briest, Shuchi Chawla, Robert Kleinberg, and S Matthew Weinberg. Pricing randomized allocations. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 585–597. SIAM, 2010.

[10] Thomas Kleine Buening, Jiarui Gan, Debmalya Mandal, and Marta Kwiatkowska. Strategyproof reinforcement learning from human feedback. *arXiv preprint arXiv:2503.09561*, 2025.

[11] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

[12] Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023.

[13] Alessandro Chiesa, Silvio Micali, and Zeyuan Allen Zhu. Mechanism design with approximate valuations. In *Proceedings of the 3rd Innovations in Theoretical Computer Science conference*, pages 34–38, 2012.

[14] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[15] Edward H Clarke. Multipart pricing of public goods. *Public choice*, pages 17–33, 1971.

[16] Vincent Conitzer and Tuomas Sandholm. Self-interested automated mechanism design and implications for optimal combinatorial auctions. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pages 132–141, 2004.

[17] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice for ai alignment: Dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

[18] Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

[19] Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[20] Michael Curry, Tuomas Sandholm, and John Dickerson. Differentiable economics for randomized affine maximizer auctions. *arXiv preprint arXiv:2202.02872*, 2022.

[21] Zvi Drezner and Horst W Hamacher. *Facility location: applications and theory*. Springer Science & Business Media, 2004.

[22] Zhijian Duan, Haoran Sun, Yurong Chen, and Xiaotie Deng. A scalable neural network for dsic affine maximizer auction design. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Zhijian Duan, Haoran Sun, Yichong Xia, Siqiang Wang, Zhilin Zhang, Chuan Yu, Jian Xu, Bo Zheng, and Xiaotie Deng. Scalable virtual valuations combinatorial auction design by combining zeroth-order and first-order optimization method. *arXiv preprint arXiv:2402.11904*, 2024.

[24] Kumar Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with llm summaries. *arXiv preprint arXiv:2404.08126*, 2024.

[25] Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. *arXiv preprint arXiv:2310.10826*, 2023.

[26] Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

[27] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

[28] Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? a systematic analysis. *arXiv preprint arXiv:2312.05488*, 2023.

[29] Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. Online advertisements with llms: Opportunities and challenges. *arXiv preprint arXiv:2311.07601*, 2023.

[30] Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. Chessgpt: Bridging policy learning and language modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

[31] Roberto Gallotta, Graham Todd, Marvin Zammit, Sam Earle, Antonios Liapis, Julian Togelius, and Georgios N Yannakakis. Large language models and games: A survey and roadmap. *arXiv preprint arXiv:2402.18659*, 2024.

[32] Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165*, 2023.

[33] Ian Gemp, Yoram Bachrach, Marc Lanctot, Roma Patel, Vibhavari Dasagi, Luke Marris, Georgios Piliouras, and Karl Tuyls. States as strings as strategies: Steering language models with game-theoretic solvers. *arXiv preprint arXiv:2402.01704*, 2024.

[34] Theodore Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631, 1973.

[35] Shangmin Guo, Haochuan Wang, Haoran Bu, Yi Ren, Dianbo Sui, Yu-Ming Shang, and Siting Lu. Large language models as rational players in competitive economics games. *arXiv preprint arXiv:2308.10032*, 2023.

[36] Shangmin Guo, Haoran Bu, Haochuan Wang, Yi Ren, Dianbo Sui, Yuming Shang, and Siting Lu. Economics arena for large language models. *arXiv preprint arXiv:2401.01735*, 2024.

[37] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

[38] Birgit Heydenreich, Rudolf Müller, Marc Uetz, and Rakesh V Vohra. Characterization of revenue equivalence. *Econometrica*, 77(1):307–316, 2009.

[39] Radosveta Ivanova-Stenzel and Timothy C Salmon. Revenue equivalence revisited. *Games and Economic Behavior*, 64(1):171–192, 2008.

[40] Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. The consensus game: Language model generation via equilibrium search. *arXiv preprint arXiv:2310.09139*, 2023.

[41] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

[42] Philippe Jehiel, Moritz Meyer-Ter-Vehn, and Benny Moldovanu. Mixed bundling auctions. *Journal of Economic Theory*, 134(1):494–512, 2007.

[43] Benjamin Laufer, Jon Kleinberg, and Hoda Heidari. Fine-tuning games: Bargaining and adaptation for general-purpose models. *arXiv preprint arXiv:2308.04399*, 2023.

[44] Anton Likhodedov and Tuomas Sandholm. Methods for boosting revenue in combinatorial auctions. In *AAAI*, pages 232–237, 2004.

[45] Nunzio Lorè and Babak Heydari. Strategic behavior of large language models: Game structure vs. contextual framing. *arXiv preprint arXiv:2309.05898*, 2023.

[46] David G Luenberger, Yinyu Ye, et al. *Linear and nonlinear programming*, volume 2. Springer, 1984.

[47] Weiyu Ma, Qirui Mi, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. *arXiv preprint arXiv:2312.11865*, 2023.

[48] James Melbourne. Strongly convex divergences. *Entropy*, 22(11):1327, 2020.

[49] Mitsunobu Miyake. On the incentive properties of multi-item auctions. *International Journal of Game Theory*, 27:1–19, 1998.

[50] Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation. *arXiv preprint arXiv:2310.08901*, 2023.

[51] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

[52] Roger B Myerson. Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, pages 61–73, 1979.

[53] Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

[54] Noam Nisan and Amir Ronen. Algorithmic mechanism design. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 129–140, 1999.

[55] Noam Nisan et al. Introduction to mechanism design (for computer scientists). *Algorithmic game theory*, 9:209–242, 2007.

[56] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[57] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[58] Susan Hesse Owen and Mark S Daskin. Strategic facility location: A review. *European journal of operational research*, 111(3):423–447, 1998.

[59] Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

[60] Alessandro Pavan, Ilya Segal, and Juuso Toikka. Dynamic mechanism design: A myersonian approach. *Econometrica*, 82(2):601–653, 2014.

[61] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

[62] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

[63] Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*, 2024.

[64] Kevin Roberts. The characterization of implementable choice rules. *Aggregation and revelation of preferences*, 12(2):321–348, 1979.

[65] Jean-Charles Rochet. A necessary and sufficient condition for rationalizability in a quasi-linear context. *Journal of mathematical Economics*, 16(2):191–200, 1987.

[66] Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

[67] Michael Saks and Lan Yu. Weak monotonicity suffices for truthfulness on convex domains. In *Proceedings of the 6th ACM conference on Electronic commerce*, pages 286–293, 2005.

[68] Tuomas Sandholm and Anton Likhodedov. Automated design of revenue-maximizing combinatorial auctions. *Operations Research*, 63(5):1000–1025, 2015.

[69] Xiao Shao, Weifu Jiang, Fei Zuo, and Mengqing Liu. Swarmbrain: Embodied agent for real-time strategy game starcraft ii via large language models. *arXiv preprint arXiv:2401.17749*, 2024.

[70] Ruizhe Shi, Yifang Chen, Yushi Hu, ALisa Liu, Noah Smith, Hannaneh Hajishirzi, and Simon Du. Decoding-time language model alignment with multiple objectives. *arXiv preprint arXiv:2406.18853*, 2024.

[71] Ermis Soumalias, Michael J Curry, and Sven Seuken. Truthful aggregation of llms with an application to online advertising. *arXiv preprint arXiv:2405.05905*, 2024.

[72] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[73] Pingzhong Tang and Tuomas Sandholm. Mixed-bundling auctions with reserve prices. In *AAMAS*, pages 729–736, 2012.

[74] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[75] William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

[76] Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*, 2024.

[77] Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. Avalon's game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint arXiv:2310.01320*, 2023.

[78] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024.

[79] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*, 2023.

[80] Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. *arXiv preprint arXiv:2310.18940*, 2023.

[81] Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024.

[82] Shun Zhang, Zhenfang Chen, Sunli Chen, Yikang Shen, Zhiqing Sun, and Chuang Gan. Improving reinforcement learning from human feedback with efficient reward model ensemble. *arXiv preprint arXiv:2401.16635*, 2024.

[83] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024.

Limitation

The main limitation of this paper is that we mainly consider the SW-Max training rules and their theoretical properties. Further study could consider more training rules and extend our model to the DPO scenario, in which each group only provides pairs of data rather than a reward model.

A Further Related Work

In this section, we review relevant research across various domains that are related to our paper, including works on RLHF with multiple reward models, multi-parameter auctions, and the intersection of game theory and LLMs.

A.1 RLHF with Multiple Reward Models

Research involving multiple reward models primarily focuses on developing algorithms to enhance practical performance. Some studies design methods simultaneously satisfying multiple preferences [11, 41, 62, 63, 70, 78, 81]. They develop more efficient algorithms to extend the Pareto frontier among different objectives [41, 62, 70, 81] and balance issues from various perspectives [11, 59, 63].

Additionally, there is a body of work that trains multiple models for a single preference and then ensembles them to improve the robustness of RLHF [18, 82], mitigate the influence of incorrect and ambiguous preferences in the dataset [76], and reduce reward hacking [26]. Unlike these approaches, our work considers how to collect misaligned preferences truthfully from different agents. As we have mentioned, these works are often assumed to be accessible to humans' actual preferences, neglecting the incentive issue for motivating rational agents to report truthfully.

A.2 Multi-parameter Auctions

Several studies have explored the properties relevant to our paper in various multi-parameter auction scenarios, such as implementability [4, 7, 16, 49, 65, 67] and payment equivalence [6, 38, 39, 60]. Another central topic in auction theory is to design mechanisms that satisfy DSIC and IR while maximizing the expected revenue for the auctioneer. Although the single-parameter scenario has been resolved by Myerson [53], the optimal auction design for multi-parameter settings remains an open question. Therefore, there is a stream of research focusing on a specific subset, affine maximizer auctions, which inherently satisfy DSIC and IR [9, 42, 44, 64, 68, 73], and proposing optimizations to enhance empirical performance [20, 22, 23]. Compared to these works, we are the first to discuss the property of payment equivalence and the revenue-maximizing solution for SW-Max training rules in the scenario of fine-tuning LLMs.

A.3 Game Theory and LLMs

In addition to the work we review in the primary related work, others have explored the intersection of game theory and large language models from different perspectives. A line of work studies other LLM-related scenarios from the algorithmic game theory perspective. Laufer et al. [43] abstracted the fine-tuning process as a bargaining game and characterized the perfect sub-game equilibria. Dubey et al. [24] proposed an auction where bidders compete to place their content within a summary generated by an LLM. Conitzer et al. [17] considered incorporating social choice theory in LLM alignment. Feizi et al. [29] explored the potential for leveraging LLMs in online advertising systems.

More broadly, some research has proposed algorithms for training LLMs inspired by concepts in game theory, such as Nash learning from human feedback [51], consensus game [40], direct Nash optimization [66], and Gemp et al. [33]. And various studies assess LLMs from a game-theoretical perspective, examining aspects such as rationality [12, 28], behavior in matrix games [2, 32, 45], and performance in strategic games like auctions [35, 36], Werewolf [79, 80], Avalon [77], Diplomacy [27, 50], card game [30] and electronic game [1, 47, 69]. There are also comprehensive surveys [31, 37, 83].

B Omitted Proofs in Section 4.1

Theorem 4.2. *In the RLHF Game with mechanism (ψ, p) that $\psi \in \Psi^{SW}$ and $p \equiv 0$, for group i, define $s_i := |\{r | r = \mathit{rm}_i(x), x \in T^*\}|$ and $\underline{\mathit{rm}_i} := \min_{\boldsymbol{x} \in T^*} \mathit{rm}_i(\boldsymbol{x})$:*

1. *If $s_i = 1$, truthfully reporting is the optimal strategy regardless of other groups' reports.*
2. *If $s_i \geq 2$ and $\underline{\mathit{rm}_i} > 0$, there is a strategy that yields strictly higher utility than truthfully reporting regardless of other groups' reports.*

Proof. If $s_i = 1$, the group gets the same utility from all training outcomes. Therefore, any strategy is optimal. We then analyze the case $s_i \geq 2$ and $\underline{\mathrm{rm}_i} > 0$ in the following. First, the optimization of ψ can be written as an equivalent constraint programming problem on the LLM_θ:

$$\begin{aligned} \arg\max_{\mathrm{LLM}_\theta} \quad & \sum_{i=1}^{n} w_i v_i(\theta; \mathrm{rm}_i) - \sum_{\boldsymbol{x} \in T^*} \mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}) f\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) \\ \text{s.t.} \quad & \sum_{\boldsymbol{x} \in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) = 1 \\ & \mathrm{LLM}_\theta(\boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in T^* \end{aligned}$$

Because of the assumption that the optimal policy satisfies $\mathrm{LLM}_\theta(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in T^*$, we can infer that the condition $\mathrm{LLM}_\theta(\boldsymbol{x}) \geq 0$, $\forall \boldsymbol{x} \in T^*$ is not active for the optimal solution. Since the convexity of the function f, by KKT condition, the necessary condition for the optimal θ^* is that there exists $\mu \in \mathbb{R}$ [46], such that

$$\sum_{i=1}^{n} w_i \frac{\partial v_i}{\partial \mathrm{LLM}_\theta(\boldsymbol{x})} - f'\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) = \mu \quad \forall \boldsymbol{x} \in T^*, \quad \sum_{\boldsymbol{x} \in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) = 1.$$

Under Definition 3.1, $\frac{\partial v_i}{\partial \mathrm{LLM}_\theta(\boldsymbol{x})} = \mathrm{rm}_i(\boldsymbol{x})$, so we have

$$\sum_{i=1}^{n} w_i \mathrm{rm}_i(\boldsymbol{x}) - f'\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) = \mu \quad \forall \boldsymbol{x} \in T^*. \tag{OPT}$$

We mainly discuss the strategies other than simply over-reporting the group size $\vec{w}$. We omit the notation $\vec{w}$ for simplicity.

Next, our main technique is to construct a report reward model $\mathrm{rm}_i' \neq \mathrm{rm}_i$ for group i such that $v_i(\psi((\mathrm{rm}_i', \overrightarrow{\mathrm{rm}}_{-i}), \theta_{\text{init}}); \mathrm{rm}_i) > v_i(\psi((\mathrm{rm}_i, \overrightarrow{\mathrm{rm}}), \theta_{\text{init}}); \mathrm{rm}_i)$ holds for all $\overrightarrow{\mathrm{rm}}_{-i}$ and θ_{init}.

The Summation Normalization Case. We first discuss the case of the reward model being normalized by summation. We take the $\boldsymbol{x}_1 \in \arg\max_{\boldsymbol{x} \in T^*} \mathrm{rm}_i(\boldsymbol{x}), \boldsymbol{x}_2 \in \arg\min_{\boldsymbol{x} \in T^*} \mathrm{rm}_i(\boldsymbol{x})$. Since $\min_{\boldsymbol{x} \in T^*} \mathrm{rm}_i(\boldsymbol{x}) > 0$, we have $\mathrm{rm}_i(\boldsymbol{x}_1) < 1$ and $\mathrm{rm}_i(\boldsymbol{x}_2) > 0$. Then we take a small $\epsilon < \min\{1 - \mathrm{rm}_i(\boldsymbol{x}_1), \mathrm{rm}_i(\boldsymbol{x}_2)\}$ and define rm_i' as:

$$\mathrm{rm}_i'(\boldsymbol{x}) = \begin{cases} \mathrm{rm}_i(\boldsymbol{x}) + \epsilon, & \boldsymbol{x} = \boldsymbol{x}_1, \\ \mathrm{rm}_i(\boldsymbol{x}) - \epsilon, & \boldsymbol{x} = \boldsymbol{x}_2 \\ \mathrm{rm}_i(\boldsymbol{x}), & \boldsymbol{x} \neq \boldsymbol{x}_1, \boldsymbol{x} \neq \boldsymbol{x}_2. \end{cases}$$

Intuitively, by reporting rm_i', group i pretends to value more for the most preferred $\boldsymbol{x}$ and less for the least preferred $\boldsymbol{x}$. Let $\theta = \psi((\mathrm{rm}_i, \overrightarrow{\mathrm{rm}}_{-i}), \theta_{\text{init}})$ and $\theta' = \psi((\mathrm{rm}_i', \overrightarrow{\mathrm{rm}}_{-i}), \theta_{\text{init}})$, we use μ and μ' to denote the variable in the necessary condition for LLM_θ and $\mathrm{LLM}_{\theta'}$, and we can derive the following results.

(a) $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) > \mathrm{LLM}_\theta(\boldsymbol{x}_1)$ and $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_2) < \mathrm{LLM}_\theta(\boldsymbol{x}_2)$. We prove the former by contradiction: if $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) \leq \mathrm{LLM}_\theta(\boldsymbol{x}_1)$, then by the convexity of f, we have

$$f'\left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) \leq f'\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right).$$

With $\mathrm{rm}_i'(\boldsymbol{x}_1) > \mathrm{rm}_i(\boldsymbol{x}_1)$, we can infer that $\mu' > \mu$. However, since for all $\boldsymbol{x} \neq \boldsymbol{x}_1$, we have $\mathrm{rm}_i'(\boldsymbol{x}) \leq \mathrm{rm}_i(\boldsymbol{x})$, to satisfy the optimal condition in (OPT), there must be for all $\boldsymbol{x} \neq \boldsymbol{x}_1$,

$$f'\left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) < f'\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right).$$

Which is equivalent to $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) < \mathrm{LLM}_{\theta}(\boldsymbol{x})$, and hence results in $\sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta'}(\boldsymbol{x}) < \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta}(\boldsymbol{x}) = 1$. The latter, $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_2) < \mathrm{LLM}_{\theta}(\boldsymbol{x}_2)$, can be proved by totally same method.

(b) The order of $\mathrm{LLM}_{\theta}(\boldsymbol{x})$ and $\mathrm{LLM}_{\theta'}(\boldsymbol{x})$ for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ is consistent. Without loss of generality, we assume there is $\boldsymbol{x}_3 \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ such that $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_3) \geq \mathrm{LLM}_{\theta}(\boldsymbol{x}_3)$. Then we have

$$f'\left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}_3)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) \geq f'\left(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x}_3)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right).$$

Then, we can infer that $\mu' \leq \mu$. For all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, to satisfy Equation (OPT), there must be

$$f'\left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) \geq f'\left(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right).$$

which is equivalent to $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) \geq \mathrm{LLM}_{\theta}(\boldsymbol{x})$. Similarly, if there is $\boldsymbol{x}_3 \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ such that $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_3) \leq \mathrm{LLM}_{\theta}(\boldsymbol{x}_3)$, then for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, there is $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) \leq \mathrm{LLM}_{\theta}(\boldsymbol{x})$.

Finally, with the results in (a) and (b), when $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) \leq \mathrm{LLM}_{\theta}(\boldsymbol{x})$ for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, the change in the utility of group i can be calculated by

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x}))\,\mathrm{rm}_i(\boldsymbol{x}) \\
=&\sum_{\boldsymbol{x}\neq\boldsymbol{x}_1,\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x}))\,\mathrm{rm}_i(\boldsymbol{x}) + (\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_{\theta}(\boldsymbol{x}_1))\,\mathrm{rm}_i(\boldsymbol{x}_1) \\
=&-\sum_{\boldsymbol{x}\neq\boldsymbol{x}_1,\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta}(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}))\,\mathrm{rm}_i(\boldsymbol{x}) + (\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_{\theta}(\boldsymbol{x}_1))\,\mathrm{rm}_i(\boldsymbol{x}_1) \\
\overset{(2)}{\geq}&-\sum_{\boldsymbol{x}\neq\boldsymbol{x}_1,\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta}(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}))\,\mathrm{rm}_i(\boldsymbol{x}_1) + (\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_{\theta}(\boldsymbol{x}_1))\,\mathrm{rm}_i(\boldsymbol{x}_1) \\
=&\mathrm{rm}_i(\boldsymbol{x}_1)\left(\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_{\theta}(\boldsymbol{x}_1) - \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1,\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta}(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}))\right) \\
=&\mathrm{rm}_i(\boldsymbol{x}_1)\sum_{\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x})) = 0.
\end{aligned}$$

When $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) \geq \mathrm{LLM}_{\theta}(\boldsymbol{x})$ for all $\boldsymbol{x} \neq \boldsymbol{x}_1, \boldsymbol{x}_2$, the change in the utility of group i can be calculated by

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x}))\,\mathrm{rm}_i(\boldsymbol{x}) \\
=&\sum_{\boldsymbol{x}\neq\boldsymbol{x}_2,\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x}))\,\mathrm{rm}_i(\boldsymbol{x}) + (\mathrm{LLM}_{\theta'}(\boldsymbol{x}_2) - \mathrm{LLM}_{\theta}(\boldsymbol{x}_2))\,\mathrm{rm}_i(\boldsymbol{x}_2) \\
=&\sum_{\boldsymbol{x}\neq\boldsymbol{x}_2,\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x}))\,\mathrm{rm}_i(\boldsymbol{x}) - (\mathrm{LLM}_{\theta}(\boldsymbol{x}_2) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}_2))\,\mathrm{rm}_i(\boldsymbol{x}_2) \\
\overset{(3)}{\geq}&\sum_{\boldsymbol{x}\neq\boldsymbol{x}_2,\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x}))\,\mathrm{rm}_i(\boldsymbol{x}_2) - (\mathrm{LLM}_{\theta}(\boldsymbol{x}_2) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}_2))\,\mathrm{rm}_i(\boldsymbol{x}_2) \\
=&\mathrm{rm}_i(\boldsymbol{x}_2)\left(\sum_{\boldsymbol{x}\neq\boldsymbol{x}_2,\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x})) - (\mathrm{LLM}_{\theta}(\boldsymbol{x}_2) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}_2))\right) \\
=&\mathrm{rm}_i(\boldsymbol{x}_2)\sum_{\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_{\theta}(\boldsymbol{x})) = 0.
\end{aligned}$$

Note that both (2) and (3) are because of $\mathrm{rm}_i(\boldsymbol{x}_1) \geq \mathrm{rm}_i(\boldsymbol{x}_2)$. And unless $\mathrm{rm}_i(\boldsymbol{x}_1) = \mathrm{rm}_i(\boldsymbol{x}_2)$, which is excluded by $s_i \geq 2$, the "$>$"s are hold.

The Maximum Normalization Case. The case of the reward model being normalized by the maximum is similar. We take the $\boldsymbol{x}_1 \in \arg\min_{\boldsymbol{x}\in T^*} \mathrm{rm}_i(\boldsymbol{x})$. Since $\min_{\boldsymbol{x}\in T^*} \mathrm{rm}_i(\boldsymbol{x}) > 0$, we have $\mathrm{rm}_i(\boldsymbol{x}_1) > 0$. Then we take a small $\epsilon < \mathrm{rm}_i(\boldsymbol{x}_1)$ and define rm'_i as:

$$\mathrm{rm}'_i(\boldsymbol{x}) = \begin{cases} \mathrm{rm}_i(\boldsymbol{x}) - \epsilon, & \boldsymbol{x} = \boldsymbol{x}_1, \\ \mathrm{rm}_i(\boldsymbol{x}), & \boldsymbol{x} \neq \boldsymbol{x}_1. \end{cases}$$

With the same technique, we first show that $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) < \mathrm{LLM}_\theta(\boldsymbol{x}_1)$ and $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) > \mathrm{LLM}_\theta(\boldsymbol{x})$ for all $\boldsymbol{x} \neq \boldsymbol{x}_1$. After that, it is easy to derive that when $s_i \geq 2$, the change in the utility of group i satisfies

$$\sum_{\boldsymbol{x}\in T^*} (\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x}))\, \mathrm{rm}_i(\boldsymbol{x}) > 0.$$

□

Lemma B.1. *When the training rule $\psi \in \Psi^{SW}$, and the divergence function f is α-strongly convex and C^2-smooth, then ψ satisfies Definition 4.7.*

Proof. As is shown in the proof of Theorem 4.2, we have two Lagrangian variables μ and μ' for $(\overrightarrow{\mathrm{rm}}, \vec{w})$ and $(\overrightarrow{\mathrm{rm}}, \vec{w})$, respectively. We have the following equations:

$$\sum_{i=1}^n w_i \mathrm{rm}_i(\boldsymbol{x}) - f'\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) = \mu, \quad \forall \boldsymbol{x} \in T^*.$$
$$\sum_{i=1}^n w'_i \mathrm{rm}'_i(\boldsymbol{x}) - f'\left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) = \mu', \quad \forall \boldsymbol{x} \in T^*.$$

Firstly, we have $|\mu' - \mu| \leq \max_{\boldsymbol{x}\in T^*} |\sum_{i=1}^n w_i \mathrm{rm}_i(\boldsymbol{x}) - \sum_{i=1}^n w'_i \mathrm{rm}'_i(\boldsymbol{x})|$. Otherwise, without loss of generality, assume that $\mu' - \mu > \max_{\boldsymbol{x}\in T^*} |\sum_{i=1}^n w_i \mathrm{rm}_i(\boldsymbol{x}) - \sum_{i=1}^n w'_i \mathrm{rm}'_i(\boldsymbol{x})|$, then we can derive that $\forall \boldsymbol{x} \in T^*$,

$$f'\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) < f'\left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right).$$

This means that $\mathrm{LLM}_\theta(\boldsymbol{x}) < \mathrm{LLM}_{\theta'}(\boldsymbol{x})$ for all $\boldsymbol{x}$, which leads the contradiction. Therefore, we have for all $\boldsymbol{x} \in T^*$

$$\begin{aligned}\left|f'\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) - f'\left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right)\right| &= \left|\sum_{i=1}^n w_i \mathrm{rm}_i(\boldsymbol{x}) - \sum_{i=1}^n w'_i \mathrm{rm}'_i(\boldsymbol{x}) + \mu' - \mu\right| \\ &\leq 2\left|\sum_{i=1}^n w_i \mathrm{rm}_i(\boldsymbol{x}) - \sum_{i=1}^n w'_i \mathrm{rm}'_i(\boldsymbol{x})\right|.\end{aligned}$$

By C^2-smoothness of f and the α-strongly convexity, we have for all $\boldsymbol{x} \in T^*$

$$\begin{aligned}|\mathrm{LLM}_\theta(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x})| &\leq \frac{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{\alpha}\left|f'\left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right) - f'\left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}\right)\right| \\ &\leq \frac{2\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{\alpha}\left|\sum_{i=1}^n w_i \mathrm{rm}_i(\boldsymbol{x}) - \sum_{i=1}^n w'_i \mathrm{rm}'_i(\boldsymbol{x})\right|.\end{aligned}$$

Therefore, for any $\epsilon > 0$, if $|\sum_{i=1}^n w_i \mathrm{rm}_i(\boldsymbol{x}) - \sum_{i=1}^n w'_i \mathrm{rm}'_i(\boldsymbol{x})| < \frac{\alpha\epsilon}{2}$, then $|\mathrm{LLM}_\theta(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x})| \leq \epsilon$. □

Theorem B.2 (Detailed version of Theorem 4.3)**.** *In the RLHF Game with mechanism (ψ, p) that $\psi \in \Psi^{SW}$ and $p \equiv 0$, when f is α-strongly convex and C^2-smooth, suppose group i has preference rm_i and group size w_i, other groups report $(\overrightarrow{rm}_{-i}, \vec{w}_{-i})$ and the initial model θ_{init}, we define*

$$t(\boldsymbol{z}) := \sum_{\boldsymbol{x}\in T^*} \frac{(rm_i(\boldsymbol{z}) - rm_i(\boldsymbol{x}))LLM_{\theta_{init}}(\boldsymbol{x})}{f''\left(\frac{LLM_\theta(\boldsymbol{x})}{LLM_{\theta_{init}}(\boldsymbol{x})}\right)},$$

in which $\theta = \psi(\overrightarrow{rm}, \vec{w}, \theta_{init})$. When $s_i \geq 2$ and $\underline{rm_i} = 0$:

1. *For the maximum normalization case, if there exist* $\boldsymbol{x}_1 \in T^*$, $t(\boldsymbol{x}_1) \neq 0$ *and* $0 < rm_i(\boldsymbol{x}_1) < 1$, *truthful reporting is not the optimal strategy.*
2. *For the summation normalization case, if there exist* $\boldsymbol{x}_1 \in T^*$, $t(\boldsymbol{x}_1) < 0$ *and* $0 < rm_i(\boldsymbol{x}_1) < 1$, *truthful reporting is not the optimal strategy.*
3. *For the summation normalization case, if there exist* $\boldsymbol{x}_1 \in T^*$, $t(\boldsymbol{x}_1) > 0$ *and we can also find* $\boldsymbol{x}_2 \in T^*$, *such that* $1 > rm_i(\boldsymbol{x}_1) \geq rm_i(\boldsymbol{x}_2) > 0$ *and* $\frac{1}{LLM_{\theta_{init}}(\boldsymbol{x}_1)} f'' \left(\frac{LLM_\theta(\boldsymbol{x}_1)}{LLM_{\theta_{init}}(\boldsymbol{x}_1)} \right) < \frac{1}{LLM_{\theta_{init}}(\boldsymbol{x}_2)} f'' \left(\frac{LLM_\theta(\boldsymbol{x}_2)}{LLM_{\theta_{init}}(\boldsymbol{x}_2)} \right)$, *truthful reporting is not the optimal strategy.*

Proof. As is shown in the proof of Theorem 4.2, the necessary condition for the solution θ is that there exists a $\mu \in \mathbb{R}$ such that

$$\sum_{i=1}^{n} w_i \mathrm{rm}_i(\boldsymbol{x}) - f' \left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} \right) = \mu \quad \forall \boldsymbol{x} \in T^*, \sum_{\boldsymbol{x} \in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) = 1.$$

And by Lemma B.1, we can also use the condition Definition 4.7.

The Maximum Normalization Case (1). Without loss of generality, we assume that there exists $\boldsymbol{x}_1$ such that $t(\boldsymbol{x}_1) > 0$, we take $0 < \epsilon < 1 - \mathrm{rm}_i(\boldsymbol{x}_1)$ to construct a report rm_i'

$$\mathrm{rm}_i'(\boldsymbol{x}) = \begin{cases} \mathrm{rm}_i(\boldsymbol{x}) + \epsilon, & \boldsymbol{x} = \boldsymbol{x}_1, \\ \mathrm{rm}_i(\boldsymbol{x}), & \boldsymbol{x} \neq \boldsymbol{x}_1. \end{cases}$$

Suppose that μ' is the Lagrangian variable for the optimal solution θ' when reporting rm_i', we can derive that

$$\mu' - \mu = w_i \epsilon \mathbb{I}_{\boldsymbol{x}=\boldsymbol{x}_1} - \left(f' \left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} \right) - f' \left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} \right) \right) \quad \forall \boldsymbol{x} \in T^*.$$

With a similar analyze in the proof of Theorem 4.2, we can induce that $\mu' > \mu$ and $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) < \mathrm{LLM}_\theta(\boldsymbol{x})$ for all $\boldsymbol{x} \neq \boldsymbol{x}_1$. By the C^2-smoothness of f, for each $\boldsymbol{x} \neq \boldsymbol{x}_1$, there exits a $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) \leq \boldsymbol{z} \leq \mathrm{LLM}_\theta(\boldsymbol{x})$ such that

$$\mu' - \mu = -f''(\frac{\boldsymbol{z}}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \left(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} \right).$$

For convenience, we let $\mathrm{LLM}_{\theta''}(\boldsymbol{x})$ refer to the corresponding $\boldsymbol{z}$ for $\boldsymbol{x}$, note that $\mathrm{LLM}_{\theta''}$ is not necessarily a distribution. We then compute the change in the group i's utility:

$$\begin{aligned}
&\sum_{\boldsymbol{x} \in T^*} \mathrm{rm}_i(\boldsymbol{x})(\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x})) \\
=&\mathrm{rm}_i(\boldsymbol{x}_1)(\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_\theta(\boldsymbol{x}_1)) + \sum_{\boldsymbol{x} \neq \boldsymbol{x}_1} \mathrm{rm}_i(\boldsymbol{x})(\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x})) \\
=&\mathrm{rm}_i(\boldsymbol{x}_1) \sum_{\boldsymbol{x} \neq \boldsymbol{x}_1} (\mathrm{LLM}_\theta(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x})) - \sum_{\boldsymbol{x} \neq \boldsymbol{x}_1} \mathrm{rm}_i(\boldsymbol{x})(\mathrm{LLM}_\theta(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x})) \\
=&\sum_{\boldsymbol{x} \neq \boldsymbol{x}_1} \frac{(\mu' - \mu)(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}))\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})}.
\end{aligned}$$

Then, we show that when the ϵ we choose is sufficiently small, the above term is positive. We define the lower bound:

$$\delta_1 := \min_{\boldsymbol{x} \in T^*} f''(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}).$$

Since function f is α-strongly convex, $\delta_1 \geq \alpha > 0$. By the C^2-smoothness of the f, there exists an $\delta_2 > 0$, such that for each θ, θ' satisfying $\max_{\boldsymbol{x}} |\mathrm{LLM}_\theta(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x})| < \delta_2$, we have

$$\max_{\boldsymbol{x} \in T^*} \left| f''(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) - f''(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \right| \leq \min\{\frac{\delta_1}{2}, \frac{\delta_1^2 t(\boldsymbol{x}_1)}{4|T^*|}\}.$$

Besides, because of the Definition 4.7, there exists δ_3, such that for each $(\vec{w}, \overrightarrow{\mathrm{rm}})$ and $(\vec{w}', \overrightarrow{\mathrm{rm}}')$ satisfying $\max_{\boldsymbol{x}\in T^*} |\sum_{i=1}^n w_i \mathrm{rm}_i(\boldsymbol{x}) - \sum_{i=1}^n w_i' \mathrm{rm}_i'(\boldsymbol{x})| \le \delta_3$, we have $\max_{\boldsymbol{x}} |\mathrm{LLM}_\theta(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x})| < \delta_2$.

Combining these, we set $\epsilon < \frac{\delta_3}{w_i}$, then it is suffice to show that

$$
\begin{aligned}
&\left| \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}))\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} - \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}))\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} \right| \\
=&\left| \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}))\left(f''(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) - f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})\right)\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \cdot f''(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} \right| \\
\le& \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{|\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x})|\left|f''(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) - f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})\right|\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \cdot f''(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} \\
<&|T^*| \cdot \frac{\delta_1^2 t(\boldsymbol{x}_1)}{4|T^*|} \cdot \frac{2}{\delta_1 \cdot \delta_1} = \frac{t(\boldsymbol{x}_1)}{2}.
\end{aligned}
$$

This means that

$$
\begin{aligned}
\sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}))\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} &> \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}))\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} - \frac{t(\boldsymbol{x}_1)}{2} \\
&= t(\boldsymbol{x}_1) - \frac{t(\boldsymbol{x}_1)}{2} = \frac{t(\boldsymbol{x}_1)}{2} > 0.
\end{aligned}
$$

Combined with $\mu' > \mu$, the proof concludes.

The Summation Normalization Case (2). Assume that there exists $\boldsymbol{x}_1$ such that $t(\boldsymbol{x}_1) < 0$, we select $\boldsymbol{x}_2 := \arg\max_{\boldsymbol{x}\in T^*} \mathrm{rm}_i(\boldsymbol{x})$ and take $0 < \epsilon < \min\{\mathrm{rm}_i(\boldsymbol{x}_1), 1 - \mathrm{rm}_i(\boldsymbol{x}_2)\}$ to construct a report rm_i'

$$
\mathrm{rm}_i'(\boldsymbol{x}) = \begin{cases} \mathrm{rm}_i(\boldsymbol{x}) - \epsilon, & \boldsymbol{x} = \boldsymbol{x}_1, \\ \mathrm{rm}_i(\boldsymbol{x}) + \epsilon, & \boldsymbol{x} = \boldsymbol{x}_2, \\ \mathrm{rm}_i(\boldsymbol{x}), & \boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}. \end{cases}
$$

Still, we use μ' to denote the Lagrangian variable for the optimal solution θ' when reporting rm_i'. Then, there are two possibilities for the relationship between μ and μ'. If $\mu \le \mu'$, by the optimal condition OPT, for all $\boldsymbol{x} \neq \boldsymbol{x}_2$, we have $\mathrm{LLM}_\theta(\boldsymbol{x}) \ge \mathrm{LLM}_{\theta'}(\boldsymbol{x})$. Since $\boldsymbol{x}_2$ has the highest reward value, such a change in the training outcome must be more preferred by the group i. Therefore, we only have to consider the case that $\mu > \mu'$. Similarly, in this case, for all $\boldsymbol{x} \neq \boldsymbol{x}_1$, we have $\mathrm{LLM}_\theta(\boldsymbol{x}) < \mathrm{LLM}_{\theta'}(\boldsymbol{x})$. By the C^2-smoothness of f, for each $\boldsymbol{x} \neq \boldsymbol{x}_1$, there exits a $\mathrm{LLM}_\theta(\boldsymbol{x}) \le \boldsymbol{z} \le \mathrm{LLM}_{\theta'}(\boldsymbol{x})$ such that

$$
\mu' - \mu = w_i \epsilon \mathbb{I}_{\boldsymbol{x}=\boldsymbol{x}_2} - f''(\frac{\boldsymbol{z}}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}).
$$

Let $\mathrm{LLM}_{\theta''}(\boldsymbol{x})$ refer to the corresponding $\boldsymbol{z}$ for $\boldsymbol{x}$, we then compute the change in the group i's utility:

$$
\begin{aligned}
&\sum_{\boldsymbol{x}\in T^*} \mathrm{rm}_i(\boldsymbol{x})(\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x})) \\
=&\mathrm{rm}_i(\boldsymbol{x}_1)(\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_\theta(\boldsymbol{x}_1)) + \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \mathrm{rm}_i(\boldsymbol{x})(\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x})) \\
=&\mathrm{rm}_i(\boldsymbol{x}_1) \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} (\mathrm{LLM}_\theta(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x})) - \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \mathrm{rm}_i(\boldsymbol{x})(\mathrm{LLM}_\theta(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x})) \\
=& \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\mu' - \mu)(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}))\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} - w_i \epsilon \frac{(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}_2))\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)}{f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x}_2)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)})} \\
\ge& \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\mu' - \mu)(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}))\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})}.
\end{aligned}
$$

With the same technique we used in the maximum normalized case (1), we can show that with sufficient small $\epsilon > 0$, the above term $\sum_{\boldsymbol{x} \neq \boldsymbol{x}_1} \frac{(\mathrm{rm}_i(\boldsymbol{x}_1) - \mathrm{rm}_i(\boldsymbol{x}))\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} < \frac{t(\boldsymbol{x}_1)}{2} < 0$. Combined with $\mu' < \mu$, the proof concludes.

The Summation Normalization Case (3). Assume that there exists $\boldsymbol{x}_1$ such that $t(\boldsymbol{x}_1) > 0$,and $\boldsymbol{x}_2$, $\mathrm{rm}_i(\boldsymbol{x}_1) \geq \mathrm{rm}_i(\boldsymbol{x}_2) > 0$, we take $0 < \epsilon < \min\{\mathrm{rm}_i(\boldsymbol{x}_2), 1 - \mathrm{rm}_i(\boldsymbol{x}_1)\}$ to construct a report rm'_i

$$\mathrm{rm}'_i(\boldsymbol{x}) = \begin{cases} \mathrm{rm}_i(\boldsymbol{x}) + \epsilon, & \boldsymbol{x} = \boldsymbol{x}_1, \\ \mathrm{rm}_i(\boldsymbol{x}) - \epsilon, & \boldsymbol{x} = \boldsymbol{x}_2, \\ \mathrm{rm}_i(\boldsymbol{x}), & \boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}. \end{cases}$$

Still, we use μ' to denote the Lagrangian variable for the optimal solution θ' when reporting rm'_i. Since we know for sure that $\mathrm{LLM}_\theta(\boldsymbol{x}_1) < \mathrm{LLM}_{\theta'}(\boldsymbol{x}_1)$ and $\mathrm{LLM}_\theta(\boldsymbol{x}_2) > \mathrm{LLM}_{\theta'}(\boldsymbol{x}_2)$, by the C^2-smoothness of f, $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_2) \leq \mathrm{LLM}_{\theta''}(\boldsymbol{x}_2) \leq \mathrm{LLM}_\theta(\boldsymbol{x}_2)$ and $\mathrm{LLM}_\theta(\boldsymbol{x}_1) \leq \mathrm{LLM}_{\theta''}(\boldsymbol{x}_1) \leq \mathrm{LLM}_{\theta'}(\boldsymbol{x}_1)$ such that

$$\begin{aligned} \mu' - \mu &= w_i\epsilon - f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)})\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_\theta(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)}, \\ \mu' - \mu &= -w_i\epsilon - f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x}_2)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)})\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}_2) - \mathrm{LLM}_\theta(\boldsymbol{x}_2)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)}. \end{aligned} \tag{2}$$

Let $\delta_1 := \min_{\boldsymbol{x}} \mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})$, by the C^2-smoothness of the f, there exists an $\delta_2 > 0$, such that for each θ, θ' satisfying $\max_{\boldsymbol{x}} |w_i\mathrm{rm}_i(\boldsymbol{x}) - w'_i\mathrm{rm}'_i(\boldsymbol{x})| < \delta_2$, we have

$$\max_{\boldsymbol{x} \in T^*} \left| f''(\frac{\mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) - f''(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \right| \leq \frac{\frac{\delta_1}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)} f'' \left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x}_2)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)} \right) - \frac{\delta_1}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)} f'' \left(\frac{\mathrm{LLM}_\theta(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)} \right)}{3}. \tag{3}$$

We take $\epsilon < \frac{\delta_2}{w_i}$ and first prove that when taking such ϵ, there is $\mu \leq \mu'$. By contradiction, if $\mu' < \mu$, then by condition Equation (OPT), for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, there is $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) > \mathrm{LLM}_\theta(\boldsymbol{x})$. Therefore, $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_\theta(\boldsymbol{x}_1) = \sum_{\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}} (\mathrm{LLM}_\theta(\boldsymbol{x}) - \mathrm{LLM}_{\theta'}(\boldsymbol{x})) + \mathrm{LLM}_\theta(\boldsymbol{x}_2) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}_2) < \mathrm{LLM}_\theta(\boldsymbol{x}_2) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}_2)$. However, by Equation (2), if $\mu' < \mu$, we get

$$f'' \left(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)} \right) \frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_\theta(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)} > f'' \left(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x}_2)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)} \right) \frac{\mathrm{LLM}_\theta(\boldsymbol{x}_2) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}_2)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)}$$

By Equation (3), we can derive that

$$f'' \left(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)} \right) \frac{1}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)} < f'' \left(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x}_2)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)} \right) \frac{1}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)},$$

and thus, we get

$$\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_\theta(\boldsymbol{x}_1) > \mathrm{LLM}_\theta(\boldsymbol{x}_2) - \mathrm{LLM}_{\theta'}(\boldsymbol{x}_2),$$

which brings the contradiction.

After proving that $\mu \leq \mu'$, we know that for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, $\mathbf{LLM}_\theta(\boldsymbol{x}) \geq \mathbf{LLM}_{\theta'}(\boldsymbol{x})$. Then, by the C^2-smoothness of f, for each $\boldsymbol{x} \neq \boldsymbol{x}_1$, there exits a $\mathrm{LLM}_{\theta'}(\boldsymbol{x}) \leq \boldsymbol{z} \leq \mathrm{LLM}_\theta(\boldsymbol{x})$ such that

$$\mu' - \mu = -w_i\epsilon\mathbb{I}_{\boldsymbol{x}=\boldsymbol{x}_2} - f''(\frac{\boldsymbol{z}}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})(\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x})}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}).$$

Let $\text{LLM}_{\theta''}(\boldsymbol{x})$ refer to the corresponding $\boldsymbol{z}$ for $\boldsymbol{x}$, we then compute the change in the group i's utility:

$$
\begin{aligned}
&\sum_{\boldsymbol{x}\in T^*} \text{rm}_i(\boldsymbol{x})(\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_{\theta}(\boldsymbol{x})) \\
=&\text{rm}_i(\boldsymbol{x}_1)(\text{LLM}_{\theta'}(\boldsymbol{x}_1) - \text{LLM}_{\theta}(\boldsymbol{x}_1)) + \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \text{rm}_i(\boldsymbol{x})(\text{LLM}_{\theta'}(\boldsymbol{x}) - \text{LLM}_{\theta}(\boldsymbol{x})) \\
=&\text{rm}_i(\boldsymbol{x}_1) \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} (\text{LLM}_{\theta}(\boldsymbol{x}) - \text{LLM}_{\theta'}(\boldsymbol{x})) - \sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \text{rm}_i(\boldsymbol{x})(\text{LLM}_{\theta}(\boldsymbol{x}) - \text{LLM}_{\theta'}(\boldsymbol{x})) \\
=&\sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\mu'-\mu)(\text{rm}_i(\boldsymbol{x}_1) - \text{rm}_i(\boldsymbol{x}))\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\text{LLM}_{\theta''}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} + w_i\epsilon\frac{(\text{rm}_i(\boldsymbol{x}_1) - \text{rm}_i(\boldsymbol{x}_2))\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)}{f''(\frac{\text{LLM}_{\theta''}(\boldsymbol{x}_2)}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)})} \\
\geq&\sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\mu'-\mu)(\text{rm}_i(\boldsymbol{x}_1) - \text{rm}_i(\boldsymbol{x}))\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\text{LLM}_{\theta''}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})}.
\end{aligned}
$$

With the same technique we used in the maximum normalized case (1), we can show that with sufficient small $\epsilon > 0$, the above term $\sum_{\boldsymbol{x}\neq\boldsymbol{x}_1} \frac{(\text{rm}_i(\boldsymbol{x}_1)-\text{rm}_i(\boldsymbol{x}))\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}{f''(\frac{\text{LLM}_{\theta''}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})})} > \frac{t(\boldsymbol{x}_1)}{2} > 0$. Combined with $\mu' < \mu$, the proof concludes. □

C Omitted Proofs in Section 4.2

Proposition 4.4. *For any $\psi \in \Psi^{SW}$, mechanism (ψ, p^{AFF}) satisfies DSIC and IR, and the payment is non-negative.*

Proof. We assume that for group i, the true reward model is rm_i, and the agent number is w_i. The reports of other groups are $(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i})$ and the initial model is θ_{init}.

(1) (ψ, p^{AFF}) satisfies DSIC.

We compare the utility between reporting (rm_i, w_i) and any other (rm'_i, w'_i). For convenience, we first simplify the notations by letting

$$
\begin{aligned}
\theta &= \psi(((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \theta_{\text{init}}), \\
\theta' &= \psi(((\text{rm}'_i, \overrightarrow{\text{rm}}_{-i}), (w'_i, \vec{w}_{-i})), \theta_{\text{init}}).
\end{aligned}
$$

The valuation of group i is the valuation for each agent multiplied by the real agent number:

$$
\begin{aligned}
v_i &= w_i v_i(\theta; \text{rm}_i), \\
v'_i &= w_i v_i(\theta'; \text{rm}_i).
\end{aligned}
$$

According to the payment rule p^{AFF}, the payment p_i for (rm_i, w_i) and p'_i for (rm'_i, w'_i) is

$$
\begin{aligned}
p_i &= \text{ASW}_{-i}(\psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) - \text{ASW}_{-i}(\theta; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \\
p'_i &= \text{ASW}_{-i}(\psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) - \text{ASW}_{-i}(\theta'; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})
\end{aligned}
$$

Therefore, we can calculate the change in the utility:

$$
\begin{aligned}
u'_i - u_i =&(v'_i - p'_i) - (v_i - p_i) \\
=&\big(w_i v_i(\theta'; \text{rm}_i) + \text{ASW}_{-i}(\theta'; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})\big) \\
&- \big(w_i v_i(\theta; \text{rm}_i) + \text{ASW}_{-i}(\theta; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})\big) \\
=&\text{ASW}((\theta'; (\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \theta_{\text{init}}) - \text{ASW}((\theta; (\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \theta_{\text{init}}) \\
\leq&0.
\end{aligned}
$$

The last inequality holds by the definition of θ

$$
\theta = \psi(((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \theta_{\text{init}}) = \arg\max_{\hat{\theta}\in\Theta} \text{ASW}((\hat{\theta}; (\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i})), \theta_{\text{init}}).
$$

Therefore, we can conclude that, for all $\overrightarrow{\text{rm}}$, $\vec{w}$, rm_i', w_i', we have

$$u_i((\overrightarrow{\text{rm}}, \vec{w}); \psi, p^{AFF}, \text{rm}_i, w_i) \geq u_i((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \psi, p^{AFF}, \text{rm}_i, w_i).$$

(2) (ψ, p^{AFF}) satisfies IR.

We reuse the notations above and denote θ_{-i} to be the optimal parameter for groups except for i, i.e. $\theta_{-i} = \psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$. When group i truthfully report its reward model rm_i and agent number w_i, the utility can be written as:

$$\begin{aligned}
u_i &= v_i - p_i \\
&= w_i v_i(\theta; \text{rm}_i) - \text{ASW}_{-i}(\theta_{-i}; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) + \text{ASW}_{-i}(\theta; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \\
&= w_i v_i(\theta; \text{rm}_i) + \text{ASW}_{-i}(\theta; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) - \text{ASW}_{-i}(\theta_{-i}; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \\
&= \text{ASW}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) - \text{ASW}_{-i}(\theta_{-i}; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \\
&\geq \text{ASW}(\theta_{-i}; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) - \text{ASW}_{-i}(\theta_{-i}; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \\
&= w_i v_i(\theta_{-i}; \text{rm}_i) + \text{ASW}_{-i}(\theta_{-i}; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) - \text{ASW}_{-i}(\theta_{-i}; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \\
&= w_i v_i(\theta_{-i}; \text{rm}_i) \geq 0.
\end{aligned}$$

Therefore, we can conclude that, for all $\overrightarrow{\text{rm}}$, $\vec{w}$, we have

$$u_i((\overrightarrow{\text{rm}}, \vec{w}); \psi, p^{AFF}, \text{rm}_i, w_i) \geq 0.$$

□

Proposition 4.6. *When $\vec{w} \equiv 1$ is public information, and the agents only report the reward models, all implementable training rules satisfy payment equivalence.*

Proof. We follow the result Theorem 1.37 in Nisan et al. [55].

Lemma C.1 (Theorem 1.37 in Nisan et al. [55])**.** *Let $\mathcal{R}_i$ be group i's preference domain. Assume that the $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_n$ are connected sets in the Euclidean space, then all implementable training rules ψ satisfy payment equivalence.*

In our paper, we assume that for all $i \in [n]$, $\mathcal{R}_i$ is the set of all non-negative and normalized $|T^*|$-dim vectors. Either in the summation normalization case or the maximum normalization case, this is a connected set in the Euclidean space. Hence, the theorem holds. □

Proposition 4.8. *SW-Max training rules with regularizations KL-divergence, $f_{\text{KL}}(x) = \lambda x \log x$, and χ^2 divergence, $f_2(x) = \lambda(x-1)^2$ ($\lambda > 0$ is a constant) are continuous.*

Proof. (1) For $f_{\text{KL}}(x) = \lambda x \log x$ (KL-divergence), since T^* is a finite set, we can rewrite the training rule ψ as an optimization problem as follows:

$$\begin{aligned}
\arg\max_{\text{LLM}_\theta} & \sum_{\boldsymbol{x} \in T^*} \left(\text{LLM}_\theta(\boldsymbol{x}) \sum_{i=1}^n w_i \text{rm}_i(\boldsymbol{x}) - \lambda \text{LLM}_\theta(\boldsymbol{x}) \log \frac{\text{LLM}_\theta(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} \right) \\
\text{s.t.} \quad & \sum_{\boldsymbol{x} \in T^*} \text{LLM}_\theta(\boldsymbol{x}) = 1 \\
& \text{LLM}_\theta(\boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in T^*.
\end{aligned}$$

Since for KL divergence, the optimal model LLM_θ must satisfy that $\text{LLM}_\theta(\boldsymbol{x}) > 0$, for all $\boldsymbol{x} \in T^*$. The necessary condition for an optimal θ is that there exists $\mu \in \mathbb{R}$, such that

$$\sum_{i=1}^n w_i \text{rm}_i(\boldsymbol{x}) - \lambda \log \frac{\text{LLM}_\theta(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} - \lambda = \mu \quad \forall \boldsymbol{x} \in T^*, \quad \sum_{\boldsymbol{x} \in T^*} \text{LLM}_\theta(\boldsymbol{x}) = 1.$$

Similarly, for the input $(\overrightarrow{\text{rm}}', \vec{w}')$, there exists $\mu' \in \mathbb{R}$, such that the optimal θ' satisfies

$$\sum_{i=1}^n w_i' \text{rm}_i'(\boldsymbol{x}) - \lambda \log \frac{\text{LLM}_{\theta'}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} - \lambda = \mu' \quad \forall \boldsymbol{x} \in T^*, \quad \sum_{\boldsymbol{x} \in T^*} \text{LLM}_{\theta'}(\boldsymbol{x}) = 1.$$

For convenience, we define $\Delta(\boldsymbol{x}) = \sum_{i=1}^{n} w_i' \mathrm{rm}_i'(\boldsymbol{x}) - \sum_{i=1}^{n} w_i \mathrm{rm}_i(\boldsymbol{x})$. Then the relationship between $\mathrm{LLM}_\theta(\boldsymbol{x})$ and $\mathrm{LLM}_{\theta'}(\boldsymbol{x})$ is given by

$$\mathrm{LLM}_{\theta'}(\boldsymbol{x}) = \mathrm{LLM}_\theta(\boldsymbol{x}) e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')}.$$

Note that we also have the condition

$$\sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta'}(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')} = 1.$$

Since $\sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')} = e^{\frac{1}{\lambda}(\mu-\mu')} \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})}$, we can infer that

$$1 = e^{\frac{1}{\lambda}(\mu-\mu')} \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})} \le e^{\frac{1}{\lambda}(\mu-\mu')} \max_{\boldsymbol{x}\in T^*} e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})},$$
$$1 = e^{\frac{1}{\lambda}(\mu-\mu')} \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})} \ge e^{\frac{1}{\lambda}(\mu-\mu')} \min_{\boldsymbol{x}\in T^*} e^{\frac{1}{\lambda}\Delta(\boldsymbol{x})}.$$

This is equivalent to

$$\min_{\boldsymbol{x}\in T^*} \Delta(\boldsymbol{x}) \le \mu' - \mu \le \max_{\boldsymbol{x}\in T^*} \Delta(\boldsymbol{x}).$$

Thus, the difference for $\mathrm{LLM}_\theta(\boldsymbol{x})$ and $\mathrm{LLM}_{\theta'}(\boldsymbol{x})$ can be bounded by

$$\begin{aligned}
|\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x})| &= \left|1 - e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')}\right| \mathrm{LLM}_\theta(\boldsymbol{x}) \\
&\le \left|1 - e^{\frac{1}{\lambda}(\Delta(\boldsymbol{x})+\mu-\mu')}\right| \\
&\le \max\{\max_{\boldsymbol{x}\in T^*} e^{\frac{2\Delta(\boldsymbol{x})}{\lambda}} - 1, \max_{\boldsymbol{x}\in T^*} 1 - e^{\frac{2\Delta(\boldsymbol{x})}{\lambda}}\}.
\end{aligned}$$

For any $\delta > 0$, when we set $\max_{\boldsymbol{x}\in T^*} |\Delta(\boldsymbol{x})| \le \min\{\frac{\lambda}{2}\log\frac{1}{1-\delta}, \frac{\lambda}{2}\log(1+\delta)\}$, we have

$$|\mathrm{LLM}_{\theta'}(\boldsymbol{x}) - \mathrm{LLM}_\theta(\boldsymbol{x})| \le \max\{\max_{\boldsymbol{x}\in T^*} e^{\frac{2\Delta(\boldsymbol{x})}{\lambda}} - 1, \max_{\boldsymbol{x}\in T^*} 1 - e^{\frac{2\Delta(\boldsymbol{x})}{\lambda}}\} \le \delta.$$

(2) For $f_2(x) = \lambda(x-1)^2$ (χ^2 divergence), since T^* is a finite set, we can rewrite the training rule ψ as an optimization problem as follows:

$$\begin{aligned}
\arg\max_{\mathrm{LLM}_\theta} & \sum_{x\in T^*} \left(\mathrm{LLM}_\theta(x) \sum_{i=1}^{n} w_i \mathrm{rm}_i(x) - \lambda \frac{(\mathrm{LLM}_\theta(x) - \mathrm{LLM}_{\theta_{\text{init}}}(x))^2}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})} \right) \\
\text{s.t.} \quad & \sum_{x\in T^*} \mathrm{LLM}_\theta(x) = 1 \\
& \mathrm{LLM}_\theta(x) \ge 0 \quad \forall x \in T^*.
\end{aligned}$$

Since we have assumed a relatively large λ so that the optimal model LLM_θ satisfies that $\mathrm{LLM}_\theta(x) > 0$, for all $x \in T^*$. The necessary condition for an optimal θ is that there exists $\mu \in \mathbb{R}$, such that

$$\sum_{i=1}^{n} w_i \mathrm{rm}_i(x) - 2\lambda \frac{\mathrm{LLM}_\theta(x) - \mathrm{LLM}_{\theta_{\text{init}}}(x)}{\mathrm{LLM}_{\theta_{\text{init}}}(x)} = \mu \quad \forall x \in T^*, \quad \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_\theta(\boldsymbol{x}) = 1.$$

Similarly, for the input $(\overrightarrow{\mathrm{rm}}', \vec{w}')$, there exists $\mu' \in \mathbb{R}$, such that the optimal θ' satisfies

$$\sum_{i=1}^{n} w_i' \mathrm{rm}_i'(x) - 2\lambda \frac{\mathrm{LLM}_{\theta'}(x) - \mathrm{LLM}_{\theta_{\text{init}}}(x)}{\mathrm{LLM}_{\theta_{\text{init}}}(x)} = \mu' \quad \forall x \in T^*, \quad \sum_{\boldsymbol{x}\in T^*} \mathrm{LLM}_{\theta'}(\boldsymbol{x}) = 1.$$

For convenience, we define $\Delta(x) = \sum_{i=1}^{n} w_i' \mathrm{rm}_i'(x) - \sum_{i=1}^{n} w_i \mathrm{rm}_i(x)$ Then the relationship between $\mathrm{LLM}_\theta(x)$ and $\mathrm{LLM}_{\theta'}(x)$ is given by

$$\mathrm{LLM}_{\theta'}(x) = \mathrm{LLM}_\theta(x) + \frac{\mathrm{LLM}_{\theta_{\text{init}}}(x)}{2\lambda}(\Delta(x) + \mu - \mu').$$

Note that we also have the condition

$$\sum_{x\in T^*} \mathrm{LLM}_{\theta'}(x) = \sum_{x\in T^*} \mathrm{LLM}_{\theta}(x) + \frac{\mathrm{LLM}_{\theta_{\text{init}}}(x)}{2\lambda}(\Delta(x) + \mu - \mu') = 1.$$

Since $\sum_{x\in T^*} \mathrm{LLM}_{\theta}(x) = 1$, we can infer that

$$\sum_{x\in T^*} \frac{\mathrm{LLM}_{\theta_{\text{init}}}(x)}{2\lambda}(\Delta(x) + \mu - \mu') = 0.$$

This is equivalent to

$$\mu' - \mu = \sum_{x\in T^*} \mathrm{LLM}_{\theta_{\text{init}}}(x)\Delta(x).$$

Thus, the difference for $\mathrm{LLM}_{\theta}(x)$ and $\mathrm{LLM}_{\theta'}(x)$ can be bounded by

$$|\mathrm{LLM}_{\theta'}(x) - \mathrm{LLM}_{\theta}(x)| = \left|\frac{\mathrm{LLM}_{\theta_{\text{init}}}(x)}{2\lambda}(\Delta(x) + \mu - \mu')\right| \le \frac{1}{\lambda}\max_{x\in T^*}|\Delta(x)|$$

For any $\delta > 0$, when we set $\max_{x\in T^*}|\Delta(x)| \le \lambda\delta$, we have

$$|\mathrm{LLM}_{\theta'}(x) - \mathrm{LLM}_{\theta}(x)| \le \frac{1}{\lambda}\max_{x\in T^*}|\Delta(x)| \le \delta.$$

□

Theorem 4.9. *An implementable training rule ψ satisfies payment equivalence if it is continuous and for $\forall i$, $\overrightarrow{rm}_{-i}$, $\vec{w}_{-i}$, θ_{init} there exists rm_i^* and θ such that $\psi((rm_i^*, \overrightarrow{rm}_{-i}), (w_i, \vec{w}_{-i}), \theta_{init}) \equiv \theta$ for all $w_i \in \mathcal{W}$. In the maximum normalization case, rm_i^* must be $\mathbb{1}$.*

Proof. We prove the equivalent version of payment equivalence: For any group i, when fixing other groups reports $(\overrightarrow{\mathrm{rm}}_{-i}, \vec{w}_{-i})$ and θ_{init}, any two payment rules p, p' that implement ψ in DSIC must satisfy that there exists a constant c, such that $p_i(\mathrm{rm}_i, w_i) - p_i'(\mathrm{rm}_i, w_i) = c$ for any rm_i and w_i. *Therefore, in the rest of the proof, we suppose fixed $(\overrightarrow{rm}_{-i}, \vec{w}_{-i})$ and θ_{init} and will omit these notations.*

Firstly, we introduce a new notation t_i to represent the combination (rm_i, w_i), whose domain is $\mathcal{R}\times\mathcal{W}$. Without specially claim, t_i is used to represented for the rm_i and w_i with the same superscript and subscript, for example, $t_i^k = (\mathrm{rm}_i^k, w_i^k)$. Then, we define the functions $l(\cdot,\cdot)$ and $V(\cdot,\cdot)$ as follows. $l(t_i', t_i)$ is the change in valuation from misreporting type t_i' to reporting type t_i truthfully. In formal,

$$l(t_i', t_i) := w_i v_i(\psi(t_i); \mathrm{rm}_i) - w_i v_i(\psi(t_i'); \mathrm{rm}_i).$$

And $V(t_i', t_i)$ refers to the smallest values of l on a finite and distinct path from t_i' to t_i

$$V(t_i', t_i) := \inf_{\substack{\text{A finite and distinct sequence} \\ [t_i^0 := t_i', t_i^1, \ldots, t_i^k, t_i^{k+1} := t_i]}} \sum_{j=0}^{k} l(t_i^j, t_i^{j+1}).$$

We prove the following lemma, which is a special case in Heydenreich et al. [38],

Lemma C.2 (Heydenreich et al. [38])**.** *In the RLHF Game, an implemented training rule ψ satisfies payment equivalence if for any agent i, and any types t_i, t_i', we have*

$$V(t_i, t_i') = -V(t_i', t_i).$$

Proof. Assume there is a mechanism (ψ, p) that satisfies DSIC. For any two types t_i, t_i' and a finite and distinct sequence $[t_i', t_i^1, \ldots, t_i^k, t_i]$, let $t_i^0 = t_i'$ and $t_i^{k+1} = t_i$, we have that

$$w_i^{j+1} v_i(\psi(t_i^{j+1}), \mathrm{rm}_i^{j+1}) - p_i(t_i^{j+1}) \ge w_i^{j+1} v_i(\psi(t_i^j), \mathrm{rm}_i^{j+1}) - p_i(t_i^j) \quad \forall 0 \le j \le k.$$

This can be rewritten as

$$w_i^{j+1} v_i(\psi(t_i^{j+1}), \mathrm{rm}_i^{j+1}) - w_i^{j+1} v_i(\psi(t_i^j), \mathrm{rm}_i^{j+1}) \ge p_i(t_i^{j+1}) - p_i(t_i^j) \quad \forall 0 \le j \le k.$$

Sum over j, we get the following inequality

$$\sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) = \sum_{j=0}^{k} w_i^{j+1} v_i(\psi(t_i^{j+1}), \text{rm}_i^{j+1}) - w_i^{j+1} v_i(\psi(t_i^j), \text{rm}_i^{j+1})$$
$$\geq \sum_{j=0}^{k} p_i(t_i^{j+1}) - p_i(t_i^j) = p(t_i) - p(t_i').$$

Since this holds for arbitrary finite and distinct sequences, we can infer that $V(t_i', t_i) \geq p(t_i) - p(t_i')$. Similarly, there is $V(t_i, t_i') \geq p(t_i') - p(t_i)$. Combining these results with $V(t_i, t_i') = -V(t_i', t_i)$, there is

$$V(t_i, t_i') = -V(t_i', t_i) \leq p(t_i') - p(t_i) \leq V(t_i, t_i'),$$

which means that $p(t_i') - p(t_i) = V(t_i, t_i')$. Note that this holds for arbitrary t_i and t_i'. Therefore, when for some t_i, the payment $p(t_i)$ is determined, then the payment for all other t_i's is determined. For example, if there are any two payment rules p and p' both implement ψ in DSIC, and we set the payment when i reports preference rm defined in Equation (5) and $w_i = 1$ as p^* and p'^* respectively, then $\forall t_i$

$$\begin{aligned} & p_i(t_i) - p_i'(t_i) \\ =& (p_i(t_i) - p^*) - (p_i'(t_i) - p'^*) + p^* - p'^* \\ =& V((\text{rm}, 1), t_i) - V((\text{rm}, 1), t_i) + p^* - p'^* \\ =& p^* - p'^*. \end{aligned}$$

Note that p^* and p'^* are not influenced by i's report, but they may vary for different $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$ and θ_{init}, which means that we can consider the term $p^* - p'^*$ as a function f on $(\overrightarrow{\text{rm}}_{-i}, \theta_{\text{init}})$. □

Then, we show that the training rule satisfying the conditions in Theorem 4.9 is sufficient for the condition stated in Lemma C.2. Firstly, we show that for any t_i, t_i', we have $V(t_i, t_i') + V(t_i', t_i) \geq 0$. By definition of the function $V(\cdot, \cdot)$, $V(t_i, t_i')$ and $V(t_i', t_i)$ correspond to the shortest path from t_i to t_i' and from t_i' to t_i respectively, which means that $V(t_i, t_i') + V(t_i', t_i)$ is the shortest weight for a cycle that goes through t_i and t_i'. Since the SW-Max training rule is implementable, we know that the weight for any cycle is non-negative by cycle monotonicity [65]. Therefore, $V(t_i, t_i') + V(t_i', t_i) \geq 0$ must be satisfied.

Then we show that for any t_i, t_i' and $\epsilon > 0$, $V(t_i, t_i') + V(t_i', t_i) \leq \epsilon$. We prove this by constructing a finite and distinct sequence $[t_i, t_i^1, \ldots, t_i^k, t_i']$ such that

$$\sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) + \sum_{j=0}^{k} l(t_i^{j+1}, t_i^j) \leq \epsilon. \tag{4}$$

This suffices for proving $V(t_i, t_i') + V(t_i', t_i) \leq \epsilon$ since $V(t_i, t_i')$ and $V(t_i', t_i)$ are the lower bound for $\sum_{j=0}^{k} l(t_i^j, t_i^{j+1})$ and $\sum_{j=0}^{k} l(t_i^{j+1}, t_i^j)$ respectively.

Initially, we rewrite the LHS of Equation (4) by using the definition of the function $l(\cdot, \cdot)$.

$$\begin{aligned} & \sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) + \sum_{j=0}^{k} l(t_i^{j+1}, t_i^j) \\ =& \sum_{j=1}^{k} \Big(w_i^{j+1} v_i(\psi(t_i^{j+1}), \text{rm}_i^{j+1}) - w_i^{j+1} v_i(\psi(t_i^j), \text{rm}_i^{j+1}) \Big) + \sum_{j=0}^{k} \Big(w_i^j v_i(\psi(t_i^j), \text{rm}_i^j) - w_i^j v_i(\psi(t_i^{j+1}), \text{rm}_i^j) \Big) \\ =& \sum_{j=0}^{k} w_i^{j+1} (\text{LLM}_{\theta^{j+1}} - \text{LLM}_{\theta^j}) \cdot \text{rm}_i^{j+1} + \sum_{j=0}^{k} w_i^j (\text{LLM}_{\theta^j} - \text{LLM}_{\theta^{j+1}}) \cdot \text{rm}_i^j \\ =& \sum_{j=0}^{k} (\text{LLM}_{\theta^{j+1}} - \text{LLM}_{\theta^j}) \cdot (w_i^{j+1} \text{rm}_i^{j+1} - w_i^j \text{rm}_i^j) \\ =& \sum_{j=0}^{k} \sum_{x \in T^*} (\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x))(w_i^{j+1} \text{rm}_i^{j+1}(x) - w_i^j \text{rm}_i^j(x)). \end{aligned}$$

In the above equations, $\theta^j = \psi(t_i^j)$ for $0 \le j \le k$ refers to the fine-tuned model when group i reports t_i^j.

By the condition, when $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$ and θ_{init} are fixed, there exits $\delta > 0$ such that if $\max_{x\in T^*} |w_i \text{rm}_i(x) - w_i' \text{rm}_i'(x)| \le \delta$, then $\max_{x\in T^*} |\text{LLM}_\theta(x) - \text{LLM}_{\theta'}(x)| \le \frac{\epsilon}{4\bar{w}}$ (in maximum normalization case, we take $\frac{\epsilon}{4\bar{w}|T^*|}$), where $\theta := \psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}); \theta_{\text{init}})$ and $\theta' := \psi((\text{rm}_i', \overrightarrow{\text{rm}}_{-i}), (w_i', \vec{w}_{-i}); \theta_{\text{init}})$.

We construct the sequence P as follows: we set $k = 2n$, $n \ge \frac{\bar{w}}{\delta} + 1$ and let $t_i^0 = t_i, t_i^{k+1} = t_i'$. For each $0 \le j \le n$,

$$w_i^j = w_i, \quad \text{rm}_i^j = \text{rm} + j(\frac{\text{rm}_i^* - \text{rm}}{n}).$$

And for each $n+1 \le j \le 2n+1$,

$$w_i^j = w_i', \quad \text{rm}_i^j = \text{rm}_i^* + (j-n-1)(\frac{\text{rm}' - \text{rm}_i^*}{n}).$$

Note that the rm_i^* is the one given by the condition in Theorem 4.9. In this construction, any rm_i^j is either an weighted average of rm and rm_i^* or rm' and rm_i^*. This ensures that all reward models in the sequence are valid (normalized by summation or maximum and non-negative). We can then divide the above equation into three parts, making the w_i the same in the first and the last parts.

$$\begin{aligned}
&\sum_{j=0}^{k} \sum_{x\in T^*} (\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x))(w_i^{j+1}\text{rm}_i^{j+1}(x) - w_i^j \text{rm}_i^j(x)) \\
=&\sum_{j=0}^{n-1} \sum_{x\in T^*} w_i(\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x))(\text{rm}_i^{j+1}(x) - \text{rm}_i^j(x)) &\text{(a)}\\
+&\sum_{x\in T^*} (\text{LLM}_{\theta^{n+1}}(x) - \text{LLM}_{\theta^n}(x))(w_i'\text{rm}_i^{n+1}(x) - w_i \text{rm}_i^n(x)) &\text{(b)}\\
+&\sum_{j=n+1}^{2n} \sum_{x\in T^*} w_i'(\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x))(\text{rm}_i^{j+1}(x) - \text{rm}_i^j(x)) &\text{(c)}
\end{aligned}$$

We first claim that (b) equals 0. This is because of the property of $\text{rm}_i^n = \text{rm}_i^{n+1} = \text{rm}_i^*$, which can induces $\text{LLM}_{\theta^n} = \text{LLM}_{\theta^{n+1}}$.

Then we turn to (a). By the construction, for any $x \in T^*$ and $0 \le j \le n-1$, $|w_i^j \text{rm}_i^j(x) - w_i^j \text{rm}_i^{j+1}(x)| \le \frac{\bar{w}}{n} \le \delta$, so that $|\text{LLM}_{\theta^j}(x) - \text{LLM}_{\theta^{j+1}}(x)| \le \frac{\epsilon}{4\bar{w}}$ holds for all x. Then we can derive that:

$$\begin{aligned}
&\sum_{j=0}^{n-1} \sum_{x\in T^*} w_i(\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x))(\text{rm}_i^{j+1}(x) - \text{rm}_i^j(x)) \\
=&\sum_{j=0}^{n-1} \sum_{x\in T^*} w_i(\text{LLM}_{\theta^{j+1}}(x) - \text{LLM}_{\theta^j}(x)) \frac{\text{rm}_i^*(x) - \text{rm}_i(x)}{n} \\
\le&\sum_{j=0}^{n-1} \sum_{x\in T^*} w_i \frac{\epsilon}{4\bar{w}} \frac{|\text{rm}_i^*(x) - \text{rm}_i(x)|}{n} \\
\le&\sum_{x\in T^*} \frac{\epsilon}{4} |\text{rm}_i^*(x) - \text{rm}_i(x)| \\
\le&\sum_{x\in T^*} \frac{\epsilon}{4} (\text{rm}_i^*(x) + \text{rm}_i(x)) \le \frac{\epsilon}{2}.
\end{aligned}$$

The case is similar to (c). By the construction, for any $x \in T^*$ and $n+1 \le j \le 2n$, $|w_i^j \text{rm}_i^j(x) - w_i^j \text{rm}_i^{j+1}(x)| \le \frac{\bar{w}}{n} \le \delta$, so that $|\text{LLM}_{\theta^j}(x) - \text{LLM}_{\theta^{j+1}}(x)| \le \frac{\epsilon}{4\bar{w}}$ holds for all x. Then we can

derive that:

$$\begin{aligned}
&\sum_{j=n+1}^{2n}\sum_{x\in T^*} w_i(\text{LLM}_{\theta^{j+1}}(x)-\text{LLM}_{\theta^j}(x))(\text{rm}_i^{j+1}(x)-\text{rm}_i^j(x))\\
&=\sum_{j=n+1}^{2n}\sum_{x\in T^*} w_i(\text{LLM}_{\theta^{j+1}}(x)-\text{LLM}_{\theta^j}(x))\frac{\text{rm}_i'(x)-\text{rm}_i^*(x)}{n}\\
&\le\sum_{j=n+1}^{2n}\sum_{x\in T^*} w_i\frac{\epsilon}{4\bar{w}}\frac{|\text{rm}_i'(x)-\text{rm}_i^*(x)|}{n}\\
&\le\sum_{x\in T^*}\frac{\epsilon}{4}|\text{rm}_i'(x)-\text{rm}_i^*(x)|\\
&\le\sum_{x\in T^*}\frac{\epsilon}{4}(\text{rm}_i'(x)+\text{rm}_i^*(x))\le\frac{\epsilon}{2}.
\end{aligned}$$

Combining the results from (a), (b), and (c), we have that under this construction,

$$\sum_{j=0}^{k} l(t_i^j,t_i^{j+1})+\sum_{j=0}^{k} l(t_i^{j+1},t_i^j)\le\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon.$$

By the arbitrariness of $\epsilon>0$, this is suffice to demonstrate that $V(t_i,t_i')+V(t_i,t_i')\le 0$.

Therefore, it is proven that

$$V(t_i,t_i')+V(t_i,t_i')=0.$$

which means that $V(t_i,t_i')=-V(t_i',t_i)$. By Lemma C.2, this is a sufficient condition for the payment equivalence of ψ. □

Corollary 4.10. *Each continuous training rule $\psi\in\Psi^{SW}$ satisfies payment equivalence.*

Proof. We construct the reward model as follows and show that this satisfies the condition in Corollary 4.10 for when the mechanism uses SW-Max training rules.

$$\text{rm}^*(x)=\begin{cases}\dfrac{1}{|T^*|} & \text{Summation Normalization Case,}\\ 1 & \text{Maximum Normalization Case.}\end{cases}\tag{5}$$

We prove this by contradiction. Assuming that there exist i, $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$, θ_{init}, w_i, w_i' such that

$$\theta:=\psi((\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i,\vec{w}_{-i}),\theta_{\text{init}})\neq\psi((\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i',\vec{w}_{-i}),\theta_{\text{init}})=:\theta'$$

We denote the further tie-breaking rule as $\succ_{\overrightarrow{\text{rm}}}$. Then, considering the optimality of θ, we have one of the following satisfied.

$$\text{ASW}(\theta;(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i,\vec{w}_{-i}),\theta_{\text{init}})>\text{ASW}(\theta';(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i,\vec{w}_{-i}),\theta_{\text{init}}),$$

or

$$\text{ASW}(\theta;(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i,\vec{w}_{-i}),\theta_{\text{init}})=\text{ASW}(\theta';(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i,\vec{w}_{-i}),\theta_{\text{init}}),\ \text{and}\ \text{LLM}_\theta\succ_{\overrightarrow{\text{rm}}}\text{LLM}_{\theta'}.$$

Note that $v_i(\theta;\text{rm}_i^*)=v_i(\theta';\text{rm}_i^*)$, and $\text{ASW}(\theta;(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i,\vec{w}_{-i}),\theta_{\text{init}})=(w_i'-w_i)v_i(\theta;\text{rm}_i^*)$ + $\text{ASW}(\theta;(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i',\vec{w}_{-i}),\theta_{\text{init}})$, we have

$$\text{ASW}(\theta;(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i',\vec{w}_{-i}),\theta_{\text{init}})>\text{ASW}(\theta';(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i',\vec{w}_{-i}),\theta_{\text{init}})$$

or

$$\text{ASW}(\theta;(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i',\vec{w}_{-i}),\theta_{\text{init}})=\text{ASW}(\theta';(\text{rm}_i^*,\overrightarrow{\text{rm}}_{-i}),(w_i',\vec{w}_{-i}),\theta_{\text{init}}),\ \text{and}\ \text{LLM}_\theta\succ_{\overrightarrow{\text{rm}}}\text{LLM}_{\theta'}.$$

Both cases contradicted the optimality of θ'. □

Theorem 4.11. *Given a continuous training rule $\psi \in \Psi^{SW}$ and a payment rule p implements it in DSIC: If p is always non-negative, it holds that for all i, $\overrightarrow{rm}$, $\vec{w}$, and θ_{init},*

$$p_i(\overrightarrow{rm}, \vec{w}, \theta_{init}) \geq p_i^{AFF}(\overrightarrow{rm}, \vec{w}, \theta_{init}).$$

If p implements ψ in IR, then for any $\epsilon > 0$ and i, there exists $\overrightarrow{rm}_{-i}$, $\vec{w}_{-i}$, and θ_{init}, such that for all rm_i and w_i,

$$p_i(\overrightarrow{rm}, \vec{w}, \theta_{init}) \leq p_i^{AFF}(\overrightarrow{rm}, \vec{w}, \theta_{init}) + \epsilon.$$

Proof. For a continuous SW-Max training rule ψ, we know that it satisfies payment equivalence. By the definition of payment equivalence, for any other payment rule p that also implements ψ in DSIC, there exists a function g_i such that

$$p_i(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) = p_i^{AFF}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) + g_i(\overrightarrow{\mathrm{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}).$$

Non-negative Payment. To ensure that $p_i(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) \geq 0$ always satisfied, we have the equivalent condition:

$$g_i(\overrightarrow{\mathrm{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \geq -\inf_{\mathrm{rm}_i, w_i} p_i^{AFF}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}).$$

However, for any $\overrightarrow{\mathrm{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}$, when we set rm_i to the uniform reward model Equation (5), we have shown in the previous proof that this will not change the training outcome regardless of the value of w_i and hence does not impact the ASW_{-i}. This means that the payment defined by the affine maximizer is exactly 0, and the RHS of the above equation will always be non-negative. Therefore, there must be $g_i \geq 0$ for all inputs, which means that for all i, $\overrightarrow{\mathrm{rm}}$, $\vec{w}$, and θ_{init}, we have $p_i(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) \geq p_i^{AFF}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}})$.

Individually Rationality. To ensure the utility of any group is not negative, we have to constrain the function g_i as follows:

$$g_i(\overrightarrow{\mathrm{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \leq \inf_{\mathrm{rm}_i, w_i} u_i^{AFF}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}),$$

where we denote u_i^{AFF} the utility of group i under the mechanism. We construct an extreme case such that the RHS can be sufficiently small. Without loss of generality, we assume that $T^* = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$. The initial model $\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1) = \epsilon$, $\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2) = 1 - \epsilon$. Group i has preference $\mathrm{rm}_i(\boldsymbol{x}_1) = 1$ and $\mathrm{rm}_i(\boldsymbol{x}_2) = 0$, and other groups have opposite preference: $\mathrm{rm}_j(\boldsymbol{x}_1) = 0$ and $\mathrm{rm}_j(\boldsymbol{x}_2) = 1$ for $j \neq i$. The group size is set to $w_k = 1$ for all $k \in [n]$.

In this case, as we have $\sum_{k=1}^n w_k \mathrm{rm}_k(\boldsymbol{x}_1) < \sum_{k=1}^n w_k \mathrm{rm}_k(\boldsymbol{x}_2)$, we can directly derived from the optimal condition Equation (OPT) that the final model satisfies that $\mathrm{LLM}_\theta(\boldsymbol{x}_1) \leq \mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)$. Since p^{AFF} is always non-negative, the utility of group i is at most $\mathrm{rm}_i(\boldsymbol{x}_1) \cdot \mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1) = \epsilon$. To ensure that p implements ψ in IR, we have to set $g_i(\overrightarrow{\mathrm{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) \leq \epsilon$ for this case. This is equivalent to $p_i(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) \leq p_i^{AFF}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}})$. □

D Omitted Proofs in Section 4.3

Lemma D.1. *For any rm, rm', if $\max_{\boldsymbol{x} \in T^*} |rm(\boldsymbol{x}) - rm'(\boldsymbol{x})| = \epsilon$, then for any model θ, we have*

$$|v(\theta; rm) - v(\theta; rm')| \leq \epsilon$$

Proof. We can derive that

$$\begin{aligned} |v(\theta; \mathrm{rm}) - v(\theta; \mathrm{rm}')| &= |\sum_{\boldsymbol{x} \in T^*} \mathrm{LLM}_\theta(\boldsymbol{x})(\mathrm{rm}(\boldsymbol{x}) - \mathrm{rm}'(\boldsymbol{x}))| \leq \sum_{\boldsymbol{x} \in T^*} \mathrm{LLM}_\theta(\boldsymbol{x})|\mathrm{rm}(\boldsymbol{x}) - \mathrm{rm}'(\boldsymbol{x})| \\ &\leq \sum_{\boldsymbol{x} \in T^*} \mathrm{LLM}_\theta(\boldsymbol{x})\epsilon = \epsilon. \end{aligned}$$

□

Lemma D.2. *Assume that for any noisy input $\overrightarrow{\widehat{rm}}$ generated from $F(\cdot|\overrightarrow{rm})$, and $i \in [n]$, there is*

$$\max_{\boldsymbol{x} \in T^*} |\widehat{rm}_i(\boldsymbol{x}) - rm_i(\boldsymbol{x})| \leq \epsilon.$$

Then for any $\psi \in \Psi^{SW}$ and $\overrightarrow{\widehat{rm}}$ generated from $F(\cdot|\overrightarrow{rm})$, the distance between the training outcome and the optimal is bounded by:

$$\mathrm{ASW}(\psi(\overrightarrow{\widehat{rm}}, \vec{w}, \theta_{init}); \overrightarrow{rm}, \vec{w}, \theta_{init}) \geq \mathrm{ASW}(\psi(\overrightarrow{rm}, \vec{w}, \theta_{init}); \overrightarrow{rm}, \vec{w}, \theta_{init}) - 2\epsilon \sum_{i=1}^{n} w_i.$$

Proof. Let $\hat{\theta} = \psi(\overrightarrow{\widehat{\mathrm{rm}}}, \vec{w}, \theta_{\text{init}})$ and $\theta = \psi(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}})$. $\hat{\theta}$ is the optimal parameter for biased input, and θ is the optimal parameter for the true input.

$$\begin{aligned}
\mathrm{ASW}(\hat{\theta}; \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) &= \sum_{i=1}^{n} w_i v_i(\hat{\theta}; \mathrm{rm}_i) - D_f(\mathrm{LLM}_{\hat{\theta}} || \mathrm{LLM}_{\theta_{\text{init}}}) \\
&\overset{(1)}{\geq} \sum_{i=1}^{n} w_i \left(v_i(\hat{\theta}; \widehat{\mathrm{rm}}_i) - \epsilon \right) - D_f(\mathrm{LLM}_{\hat{\theta}} || \mathrm{LLM}_{\theta_{\text{init}}}) \\
&= \mathrm{ASW}(\hat{\theta}; \overrightarrow{\widehat{\mathrm{rm}}}, \vec{w}, \theta_{\text{init}}) - \sum_{i=1}^{n} w_i \epsilon \\
&\overset{(2)}{\geq} \mathrm{ASW}(\theta; \overrightarrow{\widehat{\mathrm{rm}}}, \vec{w}, \theta_{\text{init}}) - \sum_{i=1}^{n} w_i \epsilon \\
&= \sum_{i=1}^{n} w_i v_i(\theta; \widehat{\mathrm{rm}}_i) - D_f(\mathrm{LLM}_{\theta} || \mathrm{LLM}_{\theta_{\text{init}}}) - \sum_{i=1}^{n} w_i \epsilon \\
&\overset{(3)}{\geq} \sum_{i=1}^{n} w_i \left(v_i(\theta; \mathrm{rm}_i) - \epsilon \right) - D_f(\mathrm{LLM}_{\theta} || \mathrm{LLM}_{\theta_{\text{init}}}) - \sum_{i=1}^{n} w_i \epsilon \\
&= \mathrm{ASW}(\theta; \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) - 2 \sum_{i=1}^{n} w_i \epsilon.
\end{aligned}$$

(1) and (3) can be directly induced by Lemma D.1, and (2) holds by the definition of $\hat{\theta}$.

$$\hat{\theta} = \psi(\overrightarrow{\widehat{\mathrm{rm}}}, \vec{w}, \theta_{\text{init}}) = \arg\max_{\theta \in \Theta} \mathrm{ASW}(\theta; \overrightarrow{\widehat{\mathrm{rm}}}, \vec{w}, \theta_{\text{init}}).$$

□

Theorem 4.12. *Assume that for any noisy input $\overrightarrow{\widehat{rm}}$ generated from $F(\cdot|\overrightarrow{rm})$, and $i \in [n]$, there is*

$$\max_{\boldsymbol{x} \in T^*} |\widehat{rm}_i(\boldsymbol{x}) - rm_i(\boldsymbol{x})| \leq \epsilon.$$

Then, with a training rule $\psi \in \Psi^{SW}$, (ψ, p^{AFF}) ensures that each group i can benefit at most $2w_i\epsilon$ from misreporting the reward model.

Proof. Recall that the calculation of payment in p^{AFF} is

$$p_i^{AFF}(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) = \mathrm{ASW}_{-i}(\psi(\overrightarrow{\mathrm{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) - \mathrm{ASW}_{-i}(\psi(\overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}); \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}).$$

Let $\vec{w} = (w_i, \vec{w}_{-i})$, the utility function can be written as:

$$\begin{aligned}
u_i((\mathrm{rm}_i', \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}; \psi, p, \mathrm{rm}_i, w_i) &= w_i v_i(\theta; \mathrm{rm}_i) - p_i^{AFF}((\mathrm{rm}_i', \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}, \theta_{\text{init}}) \\
&= w_i v_i(\theta; \mathrm{rm}_i) - \mathrm{ASW}_{-i}(\theta_{-i}; \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) + \mathrm{ASW}_{-i}(\theta; \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) \\
&= \mathrm{ASW}(\theta; \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}) - \mathrm{ASW}_{-i}(\theta_{-i}; \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}}),
\end{aligned}$$

where we define $\theta = \psi((\mathrm{rm}_i', \overrightarrow{\mathrm{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$, and $\theta_{-i} = \psi(\overrightarrow{\mathrm{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$. Note that the term $\mathrm{ASW}_{-i}(\theta_{-i}; \overrightarrow{\mathrm{rm}}, \vec{w}, \theta_{\text{init}})$ is not influenced by the change of rm_i or w_i.

Therefore, we can derive that for any $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}$, let $\theta_{-i} = \psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$:

$$
\begin{aligned}
&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}_i)} \left[u_i((\widehat{\text{rm}}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}; \psi, p, \text{rm}_i, w_i) + \text{ASW}_{-i}(\theta_{-i}; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) \right] \\
=&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}_i)} \left[\text{ASW}(\hat{\theta}; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) \right] \\
=&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}_i)} \left[w_i v_i(\hat{\theta}; \text{rm}_i) + \sum_{j \neq i} w_j v_j(\hat{\theta}; \text{rm}_j) - D_f(\text{LLM}_{\hat{\theta}} || \text{LLM}_{\theta_{\text{init}}}) \right] \\
\overset{(1)}{\geq}&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}_i)} \left[w_i v_i(\hat{\theta}; \widehat{\text{rm}}_i) + \sum_{j \neq i} w_j v_j(\hat{\theta}; \text{rm}_j) - D_f(\text{LLM}_{\hat{\theta}} || \text{LLM}_{\theta_{\text{init}}}) \right] - w_i\epsilon \\
\overset{(2)}{\geq}&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}_i)} \left[w_i v_i(\theta; \widehat{\text{rm}}_i) + \sum_{j \neq i} w_j v_j(\theta; \text{rm}_j) - D_f(\text{LLM}_{\theta} || \text{LLM}_{\theta_{\text{init}}}) \right] - w_i\epsilon \\
\overset{(3)}{\geq}&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}_i)} \left[w_i v_i(\theta; \text{rm}_i) + \sum_{j \neq i} w_j v_j(\theta; \text{rm}_j) - D_f(\text{LLM}_{\theta} || \text{LLM}_{\theta_{\text{init}}}) \right] - 2w_i\epsilon \\
\overset{(4)}{=}&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}'_i)} \left[w_i v_i(\theta; \text{rm}_i) + \sum_{j \neq i} w_j v_j(\theta; \text{rm}_j) - D_f(\text{LLM}_{\theta} || \text{LLM}_{\theta_{\text{init}}}) \right] - 2w_i\epsilon \\
\overset{(5)}{\geq}&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}'_i)} \left[w_i v_i(\hat{\theta}; \text{rm}_i) + \sum_{j \neq i} w_j v_j(\hat{\theta}; \text{rm}_j) - D_f(\text{LLM}_{\hat{\theta}} || \text{LLM}_{\theta_{\text{init}}}) \right] - 2w_i\epsilon \\
=&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}'_i)} \left[\text{ASW}(\hat{\theta}; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) \right] - 2w_i\epsilon \\
=&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}'_i)} \left[u_i((\widehat{\text{rm}}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}; \psi, p, \text{rm}_i, w_i) + \text{ASW}_{-i}(\theta_{-i}; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) \right] - 2w_i\epsilon
\end{aligned}
$$

All the $\hat{\theta}$ in the above inequalities refers to the optimal parameter for input $(\widehat{\text{rm}}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}}$, i.e. $\hat{\theta} = \psi((\widehat{\text{rm}}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$. Specifically, (1) and (3) come from the bounded distance between rm_i and $\widehat{\text{rm}}_i$ (Lemma D.1). (2) and (5) hold by the definitions: $\hat{\theta} = \psi((\widehat{\text{rm}}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$ = $\arg\max_{\theta' \in \Theta} \text{ASW}(\theta'; (\widehat{\text{rm}}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$ and $\theta = \psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$ = $\arg\max_{\theta' \in \Theta} \text{ASW}(\theta'; (\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$. And (4) holds since the inner term is irrelevant to $\widehat{\text{rm}}_i$.

Therefore, we get

$$
\begin{aligned}
&U_i((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}; \psi, p, \text{rm}_i, w_i) \\
=&\mathbb{E}_{\widehat{\overrightarrow{\text{rm}}} \sim \mathcal{F}(\cdot|(\text{rm}_i, \overrightarrow{\text{rm}}_{-i}))} \left[u_i(\widehat{\overrightarrow{\text{rm}}}, \vec{w}; \psi, p, \text{rm}_i, w_i) \right] \\
=&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}_i)} \mathbb{E}_{\widehat{\overrightarrow{\text{rm}}}_{-i} \sim \mathcal{F}_{-i}(\cdot|\overrightarrow{\text{rm}}_{-i})} \left[u_i((\widehat{\text{rm}}_i, \widehat{\overrightarrow{\text{rm}}}_{-i}), \vec{w}; \psi, p, \text{rm}_i, w_i) \right] \\
\geq&\mathbb{E}_{\widehat{\text{rm}}_i \sim \mathcal{F}_i(\cdot|\text{rm}'_i)} \mathbb{E}_{\widehat{\overrightarrow{\text{rm}}}_{-i} \sim \mathcal{F}_{-i}(\cdot|\overrightarrow{\text{rm}}_{-i})} \left[u_i((\widehat{\text{rm}}_i, \widehat{\overrightarrow{\text{rm}}}_{-i}), \vec{w}; \psi, p, \text{rm}_i, w_i) - 2w_i\epsilon \right] \\
=&\mathbb{E}_{\widehat{\overrightarrow{\text{rm}}} \sim \mathcal{F}(\cdot|(\text{rm}'_i, \overrightarrow{\text{rm}}_{-i}))} \left[u_i(\widehat{\overrightarrow{\text{rm}}}, \vec{w}; \psi, p, \text{rm}_i, w_i) - 2w_i\epsilon \right] \\
=&U_i((\text{rm}'_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}; \psi, p, \text{rm}_i, w_i) - 2w_i\epsilon..
\end{aligned}
$$

□

E Further Discussion on General Training Rules

In practice, some other training principles do not belong to SW-Max training rules, including those that maximize the Nash Social Welfare and focus more on fairness issues, like MaxMin-RLHF [11]. As an initial study on the incentive property of the RLHF Game, we primarily consider the mainstream training rules, SW-Max training rules, that aim to maximize social welfare under certain regularization.

Therefore, analyzing the properties of general forms of training rules is out of the scope of this paper. However, we also make a preliminary step for analyzing the two questions proposed in Section 4.2. The second question is partly included in Theorem 4.9, and for the implementability of a training rule, we utilize the notion of *cycle monotonicity* proposed by Rochet [65], which is a generalized version of monotonicity defined in a single-parameter scenario [53]. In the RLHF Game, we use the notation t_i to represent the combination of (rm_i, w_i) with the same superscript and subscript. We define the function $l(t_i', t_i; \vec{t}_{-i}, \theta_{\text{init}}) := w_i v_i(\psi((t_i, \vec{t}_{-i}), \theta_{\text{init}}); \text{rm}_i) - w_i v_i(\psi((t_i', \vec{t}_{-i}, \theta_{\text{init}})); \text{rm}_i)$ to measure group i's valuation gains from misreporting (t_i') to truthfully reporting (t_i) under $\vec{t}_{-i}$ and θ_{init}. The cycle monotonicity is defined based on this function:

Definition E.1 (Cycle Monotonicity)**.** The training rule ψ satisfies cycle monotonicity if for any group i, $t_i, t_i' \in \mathcal{R} \times \mathcal{W}$, any finite, distinct sequence of reward models $[t_i, t_i^1, t_i^2, \ldots, t_i^k, t_i']$ $(k \geq 0)$, and any $\vec{t}_{-i}, \theta_{\text{init}}$, defining $t_i^0 = t_i^{k+2} := t_i$ and $t_i^{k+1} := t_i'$, we have

$$\sum_{j=0}^{k+1} l(t_i^j, t_i^{j+1}; \vec{t}_{-i}, \theta_{\text{init}}) \geq 0.$$

For general training rules, cycle monotonicity is a sufficient and necessary condition for implementability.

Proposition E.2 (Rochet [65])**.** *A training rule ψ is implementable if and only if it satisfies cycle monotonicity.*

Proof. We fix the other groups' report $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$, θ_{init}, and also omit their notations for simplicity.

We first prove the necessity: if ψ is implementable, it satisfies cycle monotonicity. Since ψ is implementable, there exists p such that (ψ, p) satisfies DSIC. We use notation t_i^j to represent the combination of (rm_i^j, w_i^j). For any types $t_i, t_i' \in \mathcal{R} \times \mathcal{W}$, any finite and distinct sequence of types $[t_i, t_i^1, t_i^2, \ldots, t_i^k, t_i']$, $k \geq 0$, we let $t_i^0 = t_i^{k+2} := t_i$ and $t_i^{k+1} := t_i'$. By the property of DSIC, we have

$$w_i^{j+1} v_i(\psi(t_i^{j+1}); \text{rm}_i^{j+1}) - p_i(t_i^{j+1}) \geq w_i^{j+1} v_i(\psi(t_i^j); \text{rm}_i^{j+1}) - p_i(t_i^j) \quad \forall 0 \leq j \leq k+1.$$

By definition of the function l, this is equivalent to

$$l(t_i^j, t_i^{j+1}) \geq p_i(t_i^{j+1}) - p_i(t_i^j) \quad \forall 0 \leq j \leq k+1.$$

Sum over all j, we get

$$\sum_{j=0}^{k+1} l(t_i^j, t_i^{j+1}) \geq \sum_{j=0}^{k+1} \left(p_i(t_i^{j+1}) - p_i(t_i^j) \right) = 0.$$

By the arbitrariness of the sequence $[t_i, t_i^1, t_i^2, \ldots, t_i^k, t_i']$, this means that ψ satisfies cycle monotonicity.

Then, we prove the sufficiency: By cycle monotonicity, we have that for any finite and distinct sequence $[t_i, t_i^1, t_i^2, \ldots, t_i^k, t_i']$,

$$\sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) + l(t_i', t_i) = \sum_{j=0}^{k+1} l(t_i^j, t_i^{j+1}) \geq 0.$$

By the arbitrariness of the sequence, we can infer that

$$V(t_i, t_i') + l(t_i', t_i) \geq 0.$$

Since $l(t_i', t_i)$ is bounded, $V(t_i, t_i')$ is also finite and $V(t_i, t_i') \geq -l(t_i', t_i)$. Then, we can establish a payment rule p such that for any agent i,

$$p_i(t_i) = V(t^*, t_i).$$

where $t^* \in \mathcal{R} \times \mathcal{W}$ is a certain type.

Then, for any $t_i = (\text{rm}_i, w_i)$, we have

$$\begin{aligned}
&w_i v_i(\psi(t_i); \text{rm}_i) - p_i(t_i) \\
=&w_i v_i(\psi(t_i); \text{rm}_i) - V(t^*, t_i) \\
=&w_i v_i(\psi(t_i'); \text{rm}_i) + l(t_i', t_i) - V(t^*, t_i) \\
\overset{(2)}{\geq}&w_i v_i(\psi(t_i'), \text{rm}_i) - V(t^*, t_i') \\
=&w_i v_i(\psi(t_i'); \text{rm}_i) - p_i(t_i').
\end{aligned}$$

Note that (2) comes from the definition of V that:

$$\begin{aligned}
V(t^*, t_i) = &\inf_{\substack{\text{A finite and distinct sequence} \\ [t_i^0 := t^*, t_i^1, \ldots, t_i^k, t_i^{k+1} := t_i]}} \sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) \\
\leq &\inf_{\substack{\text{A finite and distinct sequence} \\ [t_i^0 := t^*, t_i^1, \ldots, t_i^k, t_i^{k+1} := t_i']}} \sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) + l(t_i', t_i) \\
= &\, V(t^*, t_i') + l(t_i', t_i).
\end{aligned}$$

This means that mechanism (ψ, p) satisfies DSIC, and suffices to show that ψ is implementable. □

Validating whether a training rule satisfies cycle monotonicity is a complex task. Thus, finding a more concise condition that can induce the implementability for a general training rule or a subset of training rules is a valuable further direction.

F Additional Experimental Results

Synthetic RLHF Game. We construct a synthetic RLHF Game: We set the group number to be 5 and assume the size of the outcome space to be 10. Each group's preference is first sampled from a uniform distribution $U[0, 1]^{10}$ and then normalized. The group sizes are uniformly sampled from $\{1, 2, \ldots, 10\}^{10}$.

We consider the misreporting strategy that is used to prove Theorem 4.2. Specifically, given a group's preference rm. We first find the most preferred and the least preferred outcome $\boldsymbol{x}_1 = \arg\max_{\boldsymbol{x}} \text{rm}(\boldsymbol{x})$, $\boldsymbol{x}_2 = \arg\min_{\boldsymbol{x}} \text{rm}(\boldsymbol{x})$. Then we set the reported reward model to be $\widetilde{\text{rm}}(\boldsymbol{x}_1) = \text{rm}(\boldsymbol{x}_1) + \epsilon$, $\widetilde{\text{rm}}(\boldsymbol{x}_2) = \text{rm}(\boldsymbol{x}_2) - \epsilon$, and $\widetilde{\text{rm}}(\boldsymbol{x}) = \text{rm}(\boldsymbol{x})$ for other $\boldsymbol{x}$s.

Table 1: Average changes in valuation and utility when adopting the misreporting strategy from Theorem 4.2, holding other groups' reports fixed. The parameter ϵ controls the extent of deviation from truthful reporting. As shown in the table, such a misreporting strategy brings valuation gain but decreases the utility.

Reporting Parameter ϵ	Type	0.001	0.002	0.005	0.01	0.02	0.05	0.1
ΔValuation (*1e2)	Mean	0.1073	0.2096	0.4881	0.8667	1.3674	1.7978	1.8154
	Std	< 0.0001	< 0.0001	0.0003	0.0004	0.0013	0.0026	0.0032
ΔUtility (*1e4)	Mean	-0.1064	-0.4135	-2.3696	-8.1557	-23.7334	-53.1552	-55.8977
	Std	< 0.0001	0.0001	0.0011	0.0046	0.0196	0.0573	0.1415

We let group 1 use this strategy and maintain the other group truthfully reporting. The payment is set according to the mechanism introduced in Section 4.2. We take $100,000$ samples and the average change in valuation and utility for group 1 is reported in Table 1. The result shows that such a strategy can indeed improve the valuation and is, hence, beneficial when there is no payment. However, with the introduced payment, no strategy will bring higher utility than truthfully reporting.

More Complex Preferences. The experiment setup of this part follows the Section 5. We consider two scenarios with more complex, multiple preferences.

1. We simulated scenarios from data reported in [70] (Table 6), involving three groups, each valuing helpfulness, harmlessness, and humor, respectively. Normalization and other settings follow our paper. The true group sizes and the numerical results are shown in the tables below.
2. We examined a scenario where the group's preference is a linear combination of two reward models. Specifically, group 1 values 0.2 $\times$ Helpfulness + 0.8 $\times$ Harmlessness, group 2 values 0.8 $\times$ Helpfulness + 0.2 $\times$ Harmlessness, and group 3 values Humor.

All of the above results show that truthfully reporting is among the optimal strategies under the mechanism.

Table 2: Valuation, utility, and social welfare outcomes when varying reporting parameters for Group 1, with other groups' reports held fixed ($\alpha = 1$ means truthful reporting). Group sizes are set as $(w_1, w_2, w_3) = (3, 2, 1)$. The three groups value Helpfulness, Harmlessness, and Humor, respectively. The highest value in each row is highlighted in **bold**.

Reporting Parameter α	0.2	0.5	1	1.5	2	3
Valuation	0.00	0.79	2.66	**3.00**	**3.00**	**3.00**
Utility (= Valuation-Payment)	0.00	0.44	**0.57**	0.50	0.50	0.50
Social Welfare	2.51	2.94	**3.08**	3.00	3.00	3.00

Table 3: Valuation, utility, and social welfare outcomes when varying reporting parameters for Group 1, with other groups' reports held fixed ($\alpha = 1$ means truthful reporting). Group sizes are set as $(w_1, w_2, w_3) = (4, 5, 3)$. The three groups value Helpfulness, Harmlessness, and Humor, respectively. The highest value in each row is highlighted in **bold**.

Reporting Parameter α	0.2	0.5	1	1.5	2	3
Valuation	0.00	0.00	1.05	1.05	3.54	**4.00**
Utility (= Valuation-Payment)	0.00	0.00	**0.43**	**0.43**	-1.83	-2.51
Social Welfare	6.51	6.51	**6.94**	**6.94**	4.68	4.00

Table 4: Valuation, utility, and social welfare outcomes when varying reporting parameters for Group 1, with other groups' reports held fixed ($\beta = 1$ means truthful reporting). Group sizes are set as $(w_1, w_2, w_3) = (5, 5, 2)$. The three groups value Helpfulness, Harmlessness, and Humor, respectively. The highest value in each row is highlighted in **bold**.

Reporting Parameter β	0.5	0.8	1	1.5	2	3
Valuation	0.00	0.33	1.31	4.67	**5.00**	**5.00**
Utility (= Valuation-Payment)	0.00	0.09	**0.20**	-0.72	-1.01	-1.01
Social Welfare	6.01	6.10	**6.20**	5.29	5.00	5.00

Table 5: Valuation, utility, and social welfare outcomes when varying reporting parameters for Group 1, with other groups' reports held fixed ($\beta = 1$ means truthful reporting). Group sizes are set as $(w_1, w_2, w_3) = (3, 1, 4)$. The three groups value Helpfulness, Harmlessness, and Humor, respectively. The highest value in each row is highlighted in **bold**.

Reporting Parameter β	0.5	0.8	1	1.5	2	3
Valuation	0.79	0.79	0.79	0.79	2.66	**3.00**
Utility (= Valuation-Payment)	**0.79**	**0.79**	**0.79**	**0.79**	-1.04	-1.58
Social Welfare	**5.37**	**5.37**	**5.37**	**5.37**	3.54	3.00

Table 6: Valuation, utility, and social welfare outcomes when varying reporting parameters for Group 1, with other groups' reports held fixed ($\alpha = 1$ means truthful reporting). Group sizes are set as $(w_1, w_2, w_3) = (2, 3, 1)$. The three groups value $0.8 \times$ Helpfulness $+ 0.2 \times$ Harmlessness, $0.2 \times$ Helpfulness $+ 0.8 \times$ Harmlessness, and Humor, respectively. The highest value in each row is highlighted in **bold**.

Reporting Parameter α	0.2	0.5	1	1.5	2	3
Valuation	0.53	0.53	1.03	1.51	**1.60**	**1.60**
Utility (= Valuation-Payment)	0.52	0.52	**0.61**	0.39	0.31	0.31
Social Welfare	2.92	2.92	**3.01**	2.79	2.71	2.71

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We are sure that the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The discussion is put in the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions and rigorous proofs are provided.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The full information combined with the code is provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use open-source datasets, and some data is simulated from certain distributions, which are described in the paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For the deterministic numerical simulation, there are no error bars. For others, we have provided clarification on the error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All information is provided in the README file of the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have checked.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.